Лекция 13. Откриване на аномалии

При откриването на аномалии целта е да се намерят обекти, които се различават от повечето други обекти. Често аномалните обекти се наричат с термина *крайности (outliers)*, тъй като при графичното изобразяване на данни те се намират далеч от останалите точки. Откриването на аномалии още известно и като *откриване на отклонения (deviation detection)*, тъй като аномалните обекти имат стойности на атрибути, които се отклоняват значително от очакваните или типичните атрибутни стойности или като *търсене на изключения (exception mining)*, тъй като аномалиите в определен смисъл са изключения.

Съществуват множество различни методи за откриването на аномалии, произтичащи от различни научни области, такива като статистика, машинно самообучение и извличане на закономерности от данни. Те всички се опитват да използват идеята, че един аномален обект е нещо необичайно или в някакъв смисъл несъвместим с други обекти. Макар необичайните обекти или събития са, по определение, относително редки, това не означава, че те не се срещат често в абсолютни термини. Например, едно събитие, което е „едно на хиляда" може да се среща милиони пъти, когато броят на разглеждани събития съставлява милиарди. Нашият интерес към аномални събития и обекти произлиза от факта, че подобни събития често имат необикновена важност, например, изключителните умения на спортиста могат да доведат до победа, аномалните стойности на някои експерименти могат да указват или на някой проблем в експеримента, или на наличието на някой нов феномен, който трябва да бъде изучен.

Следващите примери илюстрират приложения, в които откриването на аномалии предизвиква значителен интерес:

Откриване на измами (fraud detection). Поведението при покупки на някой, който е откраднал кредитната карта, вероятно се отличава от поведението на истинския притежател на картата. Компаниите, издаващи кредитни карти, опитват да открият крадците чрез анализ на покупателни шаблони, характеризиращи крадеца, или чрез отбелязване на отклонения в типичното поведение при покупки. Подобните подходи се използват и за откриване на други видове измами.

Откриване на неправомерни проникванията или намеси (intrusion detection). За съжаление атаките върху компютърни системи и мрежи се превърнали в едно често срещано явление. Докато някои от тях, такива, които са предназначени за разбиване или изваждане от строя на компютри или мрежи, са очевидни, други атаки, такива, като тези, направени да събират тайно информация, са много трудни за откриване. Тези проникванията могат да бъдат открити само чрез непрекъснато следене на системи или мрежи за наличието на някакво необичайно поведение.

Нарушения в екосистеми (ecosytem disturbances). В естествения свят има нетипични събития, които могат да оказват значително влияние върху хората. Примерите са урагани, наводнения, засухи, пожари, земетресения и т.н. Целта е да се предвиди вероятността за случване на подобни събития и изясняване на техните причини.

Общественото здраве (public health). В много страни болниците и медицинските клиники предоставят на национални организации различни статистики за по-

нататъшния анализ. Например, ако всички деца в някой град са ваксинирани срещу определено заболяване, то наличието на няколко случаи на това заболяване, разпределено между различни болници в града, е едно аномално събитие, което може да указва на наличието на някой проблем с програмата за ваксинация в града.

Медицина. В случая на отделен пациент наличието на необикновени симптоми или резултати от тестове може да указва на потенциални проблеми със здравето. Обаче, дали някой отделен резултат на теста е аномален, може да зависи от други характеристики на пациента, като, например, възрастта или пол. Освен това категоризацията един резултат като аномален или не има своя цена – ненужните допълнителни тестове, ако пациентът е здрав, или потенциалното увреждане на пациента, ако неговото състояние остава недиагностицирано или неизлекувано.

Макар че напоследък интересът към откриването на аномалии е бил задвижван от приложения, в които аномалиите са на фокус, исторически откриването на аномалии (и тяхното отстраняване) е било разглеждано като една техника за подобряване на анализа на типични обекти от данни. Например, едно относително малко количество от крайности може значително да изкриви средното и стандартното отклонение на някое множество от стойности или да промени множеството от клъстери, произвеждани от някой алгоритъм за клъстеризацията. По тази причина откриването на аномалии (и тяхното премахване) често се разглежда като една част от подготовката на данни.

13.1. Общи понятия и схеми за класификация на подходи

Преди да дискутираме специфични алгоритми за откриване на аномалии, ще въведем някои основни понятия. По-точно ще разгледаме 1) причини за аномалии, 2) подходи към откриване на аномалии, 3) разликите между подходите в зависимост от това, дали те използват или не информацията за класове и 4) изводи, които са общи за различни техники за откриване на аномалии.

13.1.1. Причини за аномалии

Най-често срещаните причини за аномалии са наличието на данни от различни класове, естественото отклонение в данни и грешки при измерване или събиране на данни.

Данни от различни класове. Един обект може да се различава от други обекти, т.е да бъде аномалия, тъй като той е просто принадлежи към другия вид или клас. Например, някой, осъществяващ измама с кредитни карти, принадлежи към един друг клас на ползватели на кредитни карти – клас, различен от класа на нормални ползватели. Идеята, че аномалните обекти идват от източника (класа), различаващ се от тези на други обекти, е формулиран в често-цитираното определение на крайност, дефинирано от статистика Дуглас Хоукинс (Douglas Hawkins):

> *Една крайност е наблюдението, което се отличава от други наблюдения до такава степен, че предизвиква подозрение, че е породено от съвсем друг механизъм.*

Естествено отклонение. Много от данни могат да се моделират чрез статистически разпределения като, например, нормалното (Гаусово) разпределение, където вероятността за появяване на един обект данни бързо намалява, когато разстоянието между обекта и центъра на разпределението нараства. С други думи, повечето от обекти са разположени около центъра (среден обект) и вероятността, че един обект се отличава значително от този среден обект е малка. Например, един изключително висок човек не е аномалия в смисъл, че принадлежи към един отделен клас обекти, но е аномалия в смисъл, че има една екстремна стойност на характеристиката (височина), притежавана от всички обекти.

Грешки при измерване или събиране на данни. Грешките при измерване или събиране на данни е един друг източник на аномалии. Например, едно измерване може да бъде неправилно записано поради човешка грешка, някой проблем с измервателното устройство или поради наличието на шума. Целта е да бъдат елиминирани подобни аномалии, тъй като те не предоставят никаква интересна информация, а само намаляват качеството на данни и на техния анализ.

И така, една аномалия може да бъде резултат от причините, описани по-горе, или да е предизвикана от други причини, които не сме разглеждали. Естествено е, че аномалиите в множества от данни могат да бъдат причинени от няколко източника и основната причина за всяка отделна аномалия е често неизвестна. На практика техниките за откриване на аномалии се фокусират върху намирането на обекти, които се отличават значително от повечето други обекти и самите техники не се влияят от конкретната причина за аномалиите. Така че причината за конкретната аномалия е важна само по отношение на използваното приложение.

13.1.2. Подходи към откриване на аномалии

Техники, базирани на модели. Много от техниките за откриване на аномалии първоначално построяват определен модел на данни. Аномалиите са обекти, които не пасват добре на модела. Например, един модел на разпределение на данни може да бъде създаден чрез използване на данни за оценяване на параметри на някое вероятностно разпределение. Един обект, който не пасва добре на модела, е аномалия, ако неговото съществуване не е много вероятно съгласно разглежданото разпределение. Ако моделът е едно множество от клъстери, то една аномалия е обектът, който не принадлежи достатъчно силно към нито един от клъстери. Когато се използва някой регресионен модел, то една аномалия е обектът, който относително далеч от стойността, предсказвана от модела.

Тъй като аномалните и нормалните обекти могат да се разглеждат като определящи два различни класа, за построяване на модела на тези класове могат да се използват различни класификационни техники. Естествено е, че класификационните техники могат да се ползват само при наличието на известни стойности на класификационен (целеви) атрибут, така че може да бъде построено някое обучаващо множество. Освен това, тъй като аномалиите са редки, това трябва да бъде отчитано както при избора на класификационната техника, така и при избора на мярка за оценяване на получения класификационен модел.

В някои случаи е трудно да се построи модела, защото, например, статистическото разпределение на данни е неизвестно или няма обучаващите данни. В такива ситуации могат да се използват техники, които не изискват построяване на модела. Някои от тях са разгледани по-долу.

Техники, базирани на близостта. Често е възможно да се дефинира някаква мярка на близост между обекти. По тази причина съществува множество подходи за определяне на аномалии, базирани на използване на такива мерки. Аномалните обекти са тези, които се намират далече от всички останали обекти. Много от техниките при този подход използват мерки за разстояние и, по тази причина, се наричат *техники за намиране на крайности, базирани на разстояние*. Когато данните могат да бъдат визуализирани чрез двумерни или тримерни графици, крайностите, базирани на разстояние, могат да бъдат намерени визуално като точки, отделени от повечето други точки.

Техники, базирани на плътността. Оценките за плътността на обекти могат да бъдат изчислени относително лесно, особено когато е зададена мярката за близост между обектите. Обектите, намиращи се в области с низка плътност, са относително по-отдалечени от техните съседи и могат да се разглеждат като аномалии. Един по-сложен подход отчита факта, че множество от данни може да съдържа области с различни плътности. Той класифицира една точка като крайност само, ако тя има локалната плътност, която значително по-малка от тази на повечето от нейните съседи.

13.1.3. Използване на информацията за класа

В зависимост от това, доколко е налична информацията за класа на анализираните данни, съществуващите подходи към откриването на аномалии могат да бъдат разделени на три основни класа – ненаправлявани, направлявани и полу-направлявани (semi-supervised).

Направляваните техники за откриване на аномалии изискват съществуване на обучаващото множество, съдържащо както аномални, така и нормални обекти (естествено е, че могат да съществуват повече от един нормален и един аномален клас). За решаване на проблема са подходящи класификационните техники, които решават така наречена *задача с рядко срещани класове (rare class problem)*, тъй като аномалиите се срещат относително по-рядко от нормалните обекти.

Ненаправляваните техники за откриване на аномалии се използват в случаи, когато липсва информация за класа на обекти. В такива случаи целта е да се присвои на всеки обект определен числов показател (score), който отразява степента, в която обектът може да се разглежда като аномален. Обърнете внимание, че присъствието на множество от аномалии, които са приличат една на друга, може да предизвика явлението, при което всички те ще бъдат маркирани като нормални или да имат нисък числов показател за аномалност. По тази причина, за да бъде една техника за ненаправляваното откриване на аномалии успешна, аномалиите трябва да бъдат различни една от друга, както и нормални обекти.

Полу-направляваните техники за откриване на аномалии използват данни, класифицирани като нормални, но нямат никаква информация за аномалните обекти. В тази ситуация целта е да се намери показател за аномалност на едно зададено

множество от обекти чрез използване на информация за нормалните обекти. Обърнете внимание, че в този случай наличието в анализираното множество от обекти на някакво множество от свързани крайности не влияе на оценката на крайностите (аномалии). Обаче в много от практическите ситуации е доста тежко да се намери едно малко множество от представителни нормални обекти.

13.1.4. Проблеми за решаване

При откриване на аномалии съществуват редица проблеми, на които трябва да бъде обърнато сериозно внимание:

Брой на атрибути, използвани за определяне на една аномалия. Въпросът дали един обект е аномален, базирайки се само на стойностите на един единствен атрибут, е въпросът до колко стойността на този атрибут за конкретния обект е аномална. Обаче, тъй като един обект може да има няколко атрибута, той може да има аномалните стойности за едни атрибути и нормалните – за други. Освен това, един обект може да бъде аномален дори ако нито една от атрибутните му стойности не са по-отделно аномални. Например, съвсем нормално е да има хора с височина 70 см (деца) или тежащи 130 кг, но е ненормално да има човек, който едновременно има ръст 70 см и тежи 130 кг. Едно общо определение за аномалия трябва да указва, как стойностите на няколко атрибута се използват за определяне, дали обектът е аномален или не. Това е от особена важност в случаи на данни с голяма размерност.

Глобална срещу локална перспектива. Един обект може да изглежда необичаен по отношение на всички обекти, но и не по отношение на обекти от неговото локално обкръжение. Например, един човек с ръст 215 см е необичайно висок в контекста на общата популация, но е съвсем нормален в контекста на професионални баскетболни играчи.

Степента, в която един обект е аномален. Оценката, дали един обект е аномален или не, се извършва от някои техники по двоичен начин, т.е един обект или аномален, или не. Често това не отговаря на съществуващото положение, при което някои от обектите са по-аномални от други. Следователно би било желателно да имаме някаква оценка за степента на аномалност на обектите. Тази оценка е известна като *показател (или степента) на аномалност (anomaly or outlier score).*

Идентифициране на аномалии една по една срещу всички аномалии отведнъж. При използване на някои техники аномалиите се отстраняват една по една, т.е. най-аномалният обект се отрива и се премахва, след това процесът се повтаря. При други техники множеството от аномалии се открива на един път. Проблемите при по-стъпковото откриване на аномалии са свързани с явлението, известно като *маскиране (masking)*, при което съществуването на няколко аномалии маскира присъствието на останалите. От друга страна, при едновременното откриване на аномалии е възможен *ефектът на затрупване (swamping)*, при който нормалните обекти се класифицират като анормалните. При подходи, базирани на модели, тези ефекти са възможни, тъй като аномалиите изкривяват използвания модел на данни.

Оценяване. При наличие на информацията за класове ефективността на една схема за откриване на аномалии може да бъде оценена чрез използване на традиционните мерки

за класификационното поведение. Обаче, тъй като класът на аномалиите е обикновено значително по-малък от класа на нормалните обекти, такива мерки като прецизност, припомняне, отношение на фалшиви грешки и т.н. (с които ще се запознаем по-подробно по-късно в тази лекция), са по-подходящи от колко традиционната класификационна точност. Ако информацията за класа не е налична, оценяването е по-трудно. Обаче за подходите, базирани на модели, ефективността на използвания метод за отриване на аномалии може да бъде преценена по отношение на подобрението на модела, получено след отстраняване на намерените аномалии.

Ефективност. Съществуват значителните разлики в изчислителната сложност на различни схеми за откриване на аномалии. Макар че класификационните схеми могат да изискват едни значителни ресурси за създаване на класификационния модел, те, обикновено, не са скъпи при своето прилагане. По същия начин, статистическите подходи, които създават статистическия модел, след това могат да го ползват за категоризация на обекти за едно постоянно време, не зависещо от броя на обектите, подлежащи на категоризацията. Подходите, основани на близостта, имат изчислителната сложност от порядъка $O(n^2)$, където n е броя на обектите, тъй като те обикновено използват информация, която се получава само след изчисляване на матрицата на близост. Времевата сложност може да бъде намалена в някои определени случаи, например за данни с малката размерност, чрез използване на специални структури данни и алгоритми.

13.2. Статистически подходи

Статистическите подходи са подходи, базирани на модели, т.е. моделът се създава от данните и обектите се оценяват по отношение на това, доколко добре те пасват на този модел. Повечето от статистическите подходи към откриване на аномалии (традиционно наричани крайности) се базират на построяване на определен модел за вероятностното разпределяне на данни и разглеждане на това, доколко обекти са вероятни при този модел. Тази идея е отразена в следното определение за крайност:

> *Една крайност е обектът, който има малката вероятност за съществуване по отношение на избрания вероятностен модел за разпределение на данни.*

Вероятностният модел за разпределение на данни се създава от данните чрез оценяване на параметрите на зададеното от потребителя вероятностно разпределение. Ако се предполага, че данните имат някакво Гаусово разпределение, то средното и стандартното отклонение на търсеното разпределение се оценяват чрез изчисляване на средното и стандартното отклонение на данните. След това се оценява вероятността, че всеки обект е породен от това разпределение.

Едни от най-важните въпроси, които възникват при използването на този подход, са:

- *Идентифициране на специфичното разпределение на конкретното множество от данни*. Докато повечето от данни могат да се опишат чрез сравнително неголям брой добре известни вероятности разпределения, такива като Гаусовото, на Поансон или биномиалното, често се срещат и множества от данни с нестандартни разпределения. Ако в подобни случаи е избран някакъв грешен модел, съвсем ясно е, че един обект може да бъде погрешно да идентифициран като крайност.

Например, данните могат да бъдат моделирани като отговарящи на Гаусовото разпределение, но в действителност да идват от някое друго разпределение, което има по-голяма вероятност (от Гаусовото) за стойностите, намиращи се далече от средната. Статистическите разпределения от този тип често се срещат на практика и са известни под името *разпределения с тежки опашки (heavy-tailed distributions)*.

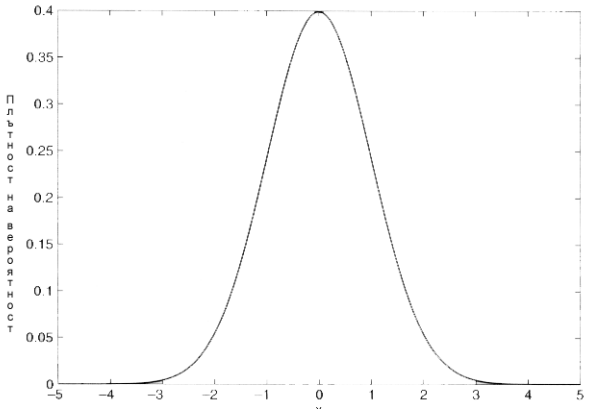
- *Броят на използвани атрибути*. Повечето от статистическите техники за откриване на крайности се прилагат към един единствен атрибут, но някои техники са приложими и за множество от атрибути.
- *Комбинации (mixtures) от разпределения.* Данните могат да се моделират като някаква комбинация от разпределения и за такива модели са разработени съответните техники за откриване на крайности. Макар че тези модели са потенциално по-изразителни, те са и по-сложни както за разбиране, така и за използване. Например, разпределенията трябва да бъдат идентифицирани преди класифициране на обекти като крайности.

13.2.1. Откриване на крайности при едномерно нормално разпределение

Гаусовото (нормално) разпределение е едно от най-често използвани разпределения в статистиката; по тази причина то ще бъде използвано в този раздел за описание на един прост подход към статистическото откриване на крайности. Разпределението има два параметъра – μ и σ, които са, съответно, средната стойност и стандартното отклонение и се означава с $N(\mu, \sigma)$. На фигура 13.1 е показана функцията на плътност на вероятност $N(0, 1)$. Съществува малък шанс един обект (стойност) от разпределението $N(0, 1)$ да се намира в опашките на разпределението. Например, вероятността че един обект се намира извън централната област (т.е. между ± 3 стандартни отклонения) е 0.0027. По-общо, ако c е някоя константа, а x е стойността на атрибута на някой обект, то вероятността, че $|x| \geq c$ намалява бързо с нарастване на c. Нека $\alpha = P(|x| \geq c)$. Таблицата 13.1 показва няколко стойности на c и съответните стойности на α за разпределението $N(0, 1)$. Обърнете внимание, че една стойност, намираща се на повечето от 4 стандартни отклонения от средната стойност, има шанс за се срещне, равен на едно на десет хиляди.

c	α за $N(0, 1)$
1.00	0.3173
1.50	0.1330
2.00	0.0455
2.50	0.0124
3.00	0.0027
3.50	0.0005
4.00	0.0001

Таблица 13.1 Двойки c и $\alpha = P(|x| \geq c)$ за $N(0, 1)$



Фиг. 13.1. Функция на плътност на вероятност при Гаусовото разпределение със средно 0 и стандартно отклонение 1

Тъй като разстоянието *c* на една стойност от центъра на разпределението *N(0, 1)* е директно свързано с вероятността на тази стойност, то тя може да се използва като основа за теста, дали обектът (стойността) е крайност или не, как това указано в следното определение:

> *Един обект с атрибутната стойност x от Гаусовото разпределение със средното 0 и стандартното отклонение 1 е крайност, ако* $|x| \geq c$, *където c е константа, избрана така, че* $P(|x| \geq c) = \alpha$.

За да се ползва тази дефиниция е необходимо да се укаже стойността на α. От гледната точка, че необичайните стойности (обекти) указват на стойността, породена от друго разпределение, α показва вероятността, че ние погрешно класифицирали някоя стойност от даденото разпределение като крайност. От гледната точка, че една крайност е рядко срещаната стойност от разпределението *N(0, 1)*, α указва на степента на рядкост.

Ако разпределение на анализиран атрибут *x* (за нормалните обекти) има Гаусовото разпределение със средното μ и стандартното отклонение σ, т.е $N(\mu, \sigma)$, то за използване на представеното по-горе определение е необходимо да трансформираме атрибутът *x* в нов атрибут *z*, който има нормалното разпределение *N(0, 1)*. Това може да се направи с трансформацията $z = (x - \mu)/\sigma$. Обаче, μ и σ обикновено не са известни и техните приблизителни стойности се намират чрез средното на извадката и стандартното отклонение на извадката. На практика, този метод работи добре при голям брой на наблюдения в извадката. При малки извадки се прилага по-сложни статистически процедури (например теста на Grubbs).

13.2.2. Крайности в многомерно нормално разпределение

За многомерното нормално Гаусово разпределение е желателно да приложим подход, сходен с този за едномерното Гаусово разпределение. В частност бих ме искали да класифицираме точки като крайности, ако те имат низка вероятност по отношение на вероятностното разпределение, оценявано чрез данни. Освен това бих ме искали да оценяваме това чрез един прост тест, например на базата на разстояние между точката и центъра на разпределението.

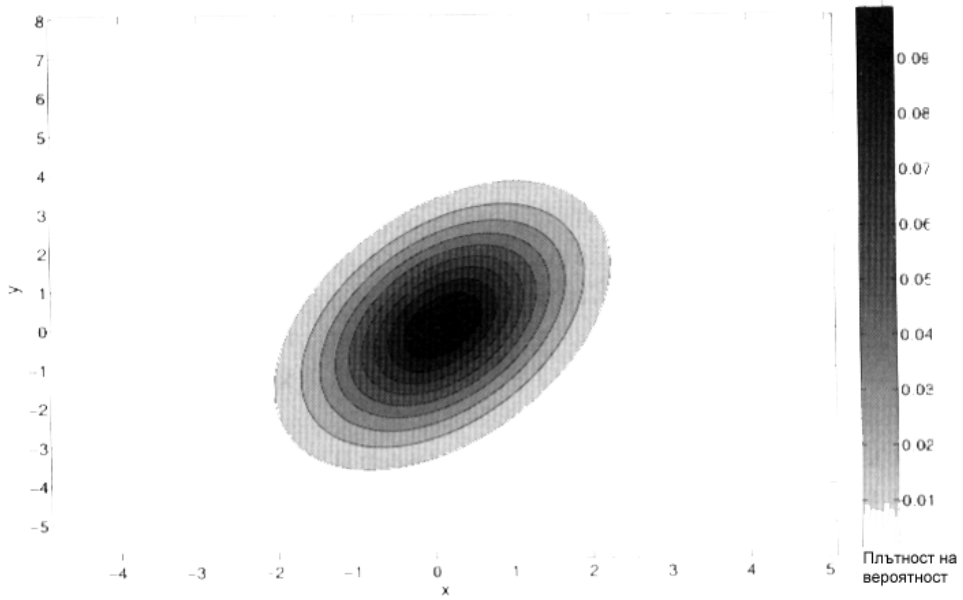
Обаче, тъй като съществува корелацията между различни атрибути, едно многомерно нормално разпределение не е симетрично по отношение на своя център. На Фиг. 13.2 е показана плътността на вероятност на едно двумерно Гаусово разпределение със средното (0, 0) и ковариационната матрица Σ.

$$\Sigma = \begin{pmatrix} 1.00 & 0.75 \\ 0.75 & 3.00 \end{pmatrix}$$

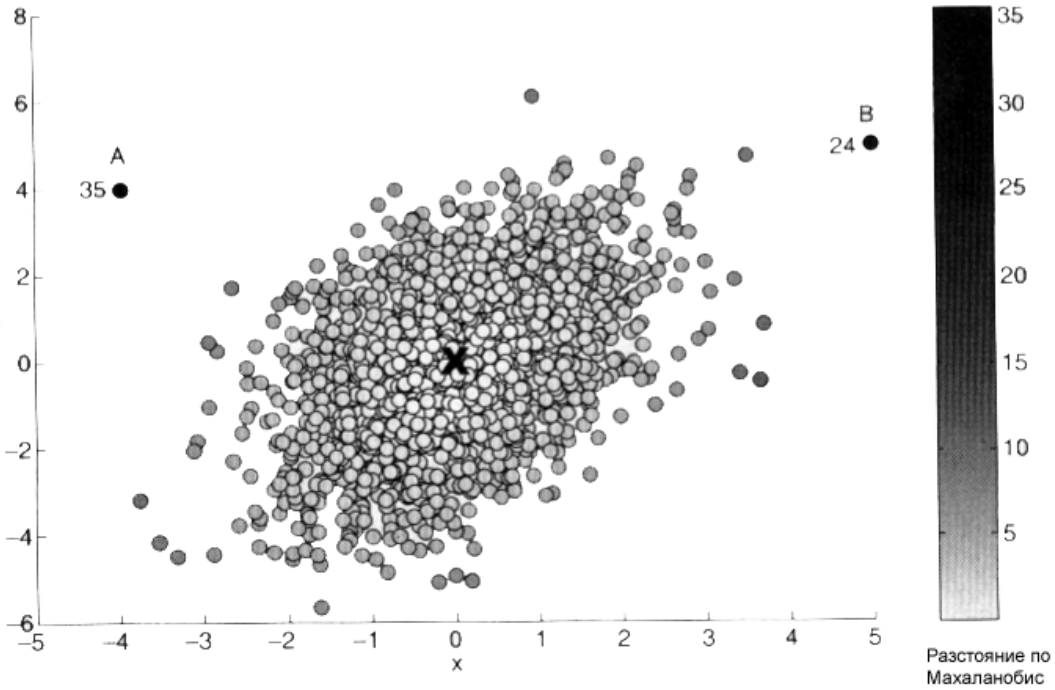
Ако за определяне на това, дали един обект е крайност или не, искаме да използваме един прост праг, ние се нуждаем от някаква мярка за разстояние, която да отчита формата на разпределението на данни. Такава мярка е разстояние по Махаланобис (виж лекция 5). Разстоянието по Махаланобис между точка **x** и средното на данни $\bar{\mathbf{x}}$ се задава от следната формула:

$$d_{Mahalanobis}(\mathbf{x}, \bar{\mathbf{x}}) = (\mathbf{x} - \bar{\mathbf{x}})\Sigma^{-1}(\mathbf{x} - \bar{\mathbf{x}})^T$$

където Σ е ковариационната матрица. Лесно е да се покаже, че разстоянието по Махаланобис от една точка до средното на анализираното разпределение е непосредствено свързано с вероятността на тази точка. В частност, разстоянието по Махаланобис е равно на логаритъма от плътността на вероятност на точката плюс константа.



Фиг. 13.2. Плътността на вероятност на точки за Гаусовото разпределение, използвано за генериране на точки от Фиг. 13.3



Фиг. 13.3. Разстояние по Махаланобис на точки до центъра на множеството от 2002 двумерни точки

Пример 13.1. Крайности при многомерно нормално разпределение. На Фиг. 13.3 е показано разстоянието по Махаланобис (от средното на разпределението) до всички

точки на едно множество от двумерни точки. Точките *A(-4, 4)* и *B(5, 5)* са крайности, които са били допълнително добавени към това множество, като тяхното разстояние по Махаланобис е указано на графика. Останалите 2000 точки от множеството са били генерирани по случаен начин, използвайки разпределението, показано на Фиг. 13.2.

И двете точки (*A* и *B*) имат голямо разстояние по Махаланобис. Обаче, макар че *A* е по-близо до центъра (голям черен знак **x** с координати (0, 0)) съгласно Евклидовото разстояние, тя е по-далеч от центъра отколкото *B* в термините на разстояние по Махаланобис, тъй като разстоянието по Махаланобис отчита формата на разпределението. В частност точката *B* има Евклидовото разстояние, равно на $5\sqrt{3}$ и разстоянието по Махаланобис равно на 24, докато точката *A* има Евклидовото разстояние $4\sqrt{2}$ и разстоянието по Махаланобис равно на 35.

13.2.3. Използване на комбинация от модели за откриване на аномалии

При този подход данните се моделират като една комбинация от два разпределения – едно за нормални данни, а друго – за крайности. Целта е да бъдат оценени параметрите на тези разпределения. В този раздел ще разгледаме един прост подход, при който в начало всички обекти се поместват в множеството от нормални обекти, а множеството от аномални обекти се разглежда като празно. След това една итеративна процедура прехвърля обекти от множеството на обикновените точки в множеството на аномалиите до тогава, докато при това прехвърляне се увеличава общата вероятност на данни.

Да предположим, че множеството данни *D* съдържа обекти от една комбинация на два разпределения: *M* – разпределението на болшинство от (нормални) обекти, и *A* – разпределението на аномални обекти. Общото вероятностно разпределение на данни може да се запише във вида:

$$D(\mathbf{x}) = (1-\lambda)M(\mathbf{x}) + \lambda A(\mathbf{x}),$$

където **x** е някой обект, а λ е някакво число между 0 и 1, задаващо каква част от всички данни се очаква да бъдат крайности. Разпределението *M* се оценява от данните, докато разпределението *A* често се приема да е равномерно. Нека M_t и A_t са, съответно, множества от нормални и аномални обекти в момент *t*. В начало $t = 0$, $M_0 = D$, $A_0 = \varnothing$. В някой момент *t* вероятността и log от вероятността на цялото множество от данни *D* се задава от следните два уравнения (предполага се, че всички точки в *D* са независими една от друга и че е изпълнено следното опростяващо предположение: вероятността е 0 при следните две ситуации – 1) когато един в обект в *A* е нормален обект и 2) когато един обект в *M* е крайност):

$$L_t(D) = \prod_{x_i \in D} P_D(\mathbf{x}_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(\mathbf{x}_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(\mathbf{x}_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(\mathbf{x}_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(\mathbf{x}_i)$$

Тъй като броят на нормални обекти е по-голям в сравнение с броя на аномални обекти, то разпределението на нормалните обекти може да не се промени много при преместването на един обект в множеството от аномални обекти. В този случай приносът на всеки нормален обект към общата вероятност на нормалните обекти ще остане относително постоянен. Освен това, ако за аномалиите се предполага едно равномерно разпределение, то всеки обект, преместен в множеството от аномалиите, добавя един постоянен принос към общата вероятност за аномалиите. Следователно, цялата промяна в общата вероятност на данни при преместването на един обект в множеството от аномалиите е приблизително равна на вероятността на този обект при равномерното вероятностно разпределение (претеглено по λ) минус вероятността на този обект при вероятностното разпределение на нормални данни (претеглено по 1 - λ). Следователно, множеството от аномалиите ще има тенденция да съдържа тези обекти, които имат значително по-голяма вероятност при едно равномерно разпределение, от колко при разпределението на нормални обекти (виж алгоритъм 13.1).

Алгоритъм 13.1. Вероятностно откриване на крайности

1. Инициализация: В момент $t = 0$ нека M_t съдържа всички обекти D, а $A_t = \varnothing$.
   Нека $LL_t(D) = LL(M_t) + LL(A_t)$ е *log* на вероятността на всички обекти.
2. **for** всяка точка **x**, принадлежаща към M_t **do**
3. Премести **x** от M_t в A_t и създай новите множества M_{t+1} и A_{t+1}
4. Изчисли новия log на вероятността на D: $LL_{t+1}(D) = LL(M_{t+1}) + LL(A_{t+1})$
5. Изчисли разликата $\Delta = LL_t(D) - LL_{t+1}(D)$
6. **if** $\Delta > c$, където c е някакъв праг, **then**
7. **x** се класифицира като аномалия, т.е. множествата M_{t+1} и A_{t+1} остават непроменени и стават текущи множества на нормални и аномални обекти.
8. **end if**
9. **end for**

Като цяло, статистическите подходи към откриване на крайности се базират върху силни теоретични основи и са построени чрез използване на стандартни статистически техники, такива като оценяване на параметрите на вероятностни разпределения. Когато има достатъчно знания за данни и типове тестове, които трябва да се приложат, тези тестове могат да бъдат много ефективни. Има много широк спектър от статистическите тестове за крайности, разработени за един единствен атрибут. По-малък избор се предлага за многомерни данни, като тези тестове могат да имат доста слабо поведение при данни с голяма размерност.

13.3. Класификационни подходи

Класификационните подходи са подходи, базирани на модели - моделът се създава от данни, които предварително класифицирани като нормални и аномални. Новите обектите се класифицират на базата на предсказването, получено чрез прилагане към тях на изградения класификационен модел. Тъй като в повечето от случаи броят на аномалиите е значително по-малък от този на нормалните обекти, откриването на аномалии може да се разглежда като един частен случай на така наречена *задача за небалансирани класове (class imbalance problem)*, която налага някои допълнителни изисквания към съществуващите алгоритми, използвани за построяване на класификационни модели.

Така мярката за точност, която интензивно се ползва за сравняване на поведение на традиционните класификатори, може да се окаже неподходяща за оценяване на модели, изведени от небалансирани множества данни. Например, ако 1% от транзакциите на кредитни карти са измамни, то един модел, който предсказва всяка транзакция като легитимна, ще има точност 99%, макар че той изобщо не може да открие никакви измами. Освен това, може да възникне необходимостта да бъдат променени и някои мерки, определящи начина за изграждане на класификационни модели (например, информационната печалба при научаване на класификационни дървета), за да се концентрират върху рядко срещани класове.

13.3.1. Алтернативни метрики

При двоичната класификация рядко срещания клас често се означава като положителен, докато често срещания клас – като отрицателен. Броят на примери, предсказани некоректно от един класификационен модел, може да бъде резюмиран чрез *матрицата на несъответствия (confusion matrix):*

		Предсказан клас	
		+	-
Действителен клас	+	f_{++}(TP)	f_{+-}(FN)
	-	f_{-+}(FP)	f_{--}(TP)

където:

- ***Действително положителни*** (TP) или f_{++} - броят на *положителни* примери, *правилно предсказани* от класификационния модел
- ***Фалшиво отрицателни*** (FN) или f_{+-} - броят на *положителни* примери, които са били *погрешно предсказани като отрицателни* от класификационния модел.
- ***Фалшиво положителни*** (FP) или f_{-+} - броят на *отрицателни* примери, които са били *погрешно предсказани като положителни* от класификационния модел.
- ***Действително отрицателни*** (TN) или f_{--} - броят на *отрицателни* примери, *правилно предсказани* от класификационния модел.

Данните от матрицата на несъответствия могат бъдат представени и във вид на проценти:

- *Процентът на действително положителни примери (TPR)* или *чувствителност* се определя като пропорция на положителни примери правилно предсказани от модела:
  $TPR = TP/(TP + FN)$
- *Процентът на действително отрицателни примери (TNR)* или *специфичност* се определя като пропорция на отрицателни примери правилно предсказани от модела:
  $TNR = TN/(TN + FP)$
- *Процентът на фалшиво положителни примери (FPR)* е пропорцията на отрицателни примери, предсказани като положителни:
  $FPR = FP/(TN + FP)$
- *Процентът на фалшиво отрицателни примери (FNR)* е пропорцията на положителни примери, предсказани като отрицателни от модела:
  $FNR = FN/(TP + FN)$

В приложения, при които успешното откриване на един от класове се разглежда като по-важно от откриване на други класове (какъвто е случай на откриване на аномалии), широко се използват следните две метрики – *припомняне (recall – r)* и *прецизност (precision – p)*:

$$p = \frac{TP}{TP + FP}, \qquad r = \frac{TP}{TP + FN}$$

Прецизността определя пропорцията на примери, които са действително положителни, измежду всички примери, предсказани от класификатора като положителни, докато припомнянето определя пропорцията на положителни примери, правилно предсказани от класификатора. Колко по-висока е прецизността, толкова е по-малък броят на фалшиво положителни примери, предсказвани от класификатора. От друга страна, класификаторите с голямата стойност на припомняне неправилно класифицират като отрицателни само много малък брой положителни примери.

Често е възможно да се създадат базови модели, които максимизират една от тези две метрики, но не и другата. Създаването на модели, максимизиращи както прецизност, така и припомняне, е основно предизвикателство пред класификационните алгоритми.

Прецизността и припомнянето могат да бъдат обединени в една друга метрика, известна като мярка F_1:

$$F_1 = \frac{2rp}{r + p} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

По принцип, F_1 представлява хармоничното средно между припомнянето и прецизност, т.е.

$$F_1 = \frac{2}{\frac{1}{r} + \frac{1}{p}}$$

Хармоничното средно на две числа е по-близо до по-малкото от числата. Следователно, високата стойност на мярката F_1 осигурява, че и двете характеристики – прецизност и припомняне са достатъчно големи.

Една по-обща мярка за баланса между прецизността и припомняне е мярката F_β:

$$F_\beta = \frac{(\beta^2 + 1)rp}{r + \beta^2 p} = \frac{(\beta^2 + 1) \times TP}{(\beta^2 + 1)TP + \beta^2 FP + FN}$$

И прецизността и припомнянето са специални случаи на F_β при $\beta = 0$ и $\beta = \infty$ съответно. Ниските стойности на β правят F_β по-близо до прецизността, а високите стойности – по-близо до припомняне.

13.3.2. Обучение с отчитане на цена

Една *ценова матрица (cost matrix)* кодира „наказания" за неправилното класифициране на един обект, принадлежащ към определен клас, като принадлежащ към някой друг

клас. Нека *C(i, j)* означава цената за предсказване на един обект от класа *i* като принадлежащ към класа *j*. Използвайки това означение *(C+, -)* е цената за направяне една фалшиво отрицателна грешка, докато *C(-, +)* е цената за генериране на едно фалшиво предупреждение (фалшиво положителна грешка). Отрицателните стойности в ценовата матрица представляват награда за правилната класификация. При зададена колекция от *N* тестови примера, *общата цена* на един класификационен модел *M* се изчислява като:

$$C(M) = TP*C(+, +) + FP*C(-, +) + FN* C(+, -) + TN*C(-, -)$$

Ако ценовата матрица съдържа само стойности 0/1, т.е $C(+, +) = C(-, -) = 0$ и $(C+, -) = C(-, +) = 1$, то общата цена на модела е еквивалентна на броя на направените грешки при класификация:

$$C(M) = 0*(TP + TN) + 1*(FP + FN) = N*Err,$$

където *Err* – процентът на допуснати класификационни грешки.

Пример 13.2. Да разгледаме ценовата матрица, представена в Таблица 13.2 – цената на една фалшиво отрицателна грешка е сто пъти по-голяма от цената на една фалшива тревога. С други думи, неуспехът при откриване на един положителен пример (например, откриване на една транзакция с откраднатата кредитна карта) е еквивалентен на ефекта от издаване на сто фалшиви тревоги (например, недопускане на 100 подозрителни транзакции с кредитни карти, принадлежащи на легални собственици). Имайки два класификационни модела, описани с матриците на несъответствия, показани в Таблица 14.3, общата цена на всеки модел е равна на:

$$C(M_1) = 150*(-1) + 60*1 + 40*100 = 3910$$
$$C(M_2) = 250*(-1) + 5*1 + 45*100 = 4255$$

		Предсказан клас	
		Клас = +	Клас = -
Действителен клас	Клас = +	-1	100
	Клас = -	1	0

Таблица 13.2. Ценовата матрица за примера 13.2

Модел M_1		***Предсказан клас***	
		Клас +	Клас -
Действителен клас	Клас +	150	40
	Клас -	60	250

Модел M_2		***Предсказан клас***	
		Клас+	Клас -
Действителен клас	Клас +	250	45
	Клас -	5	200

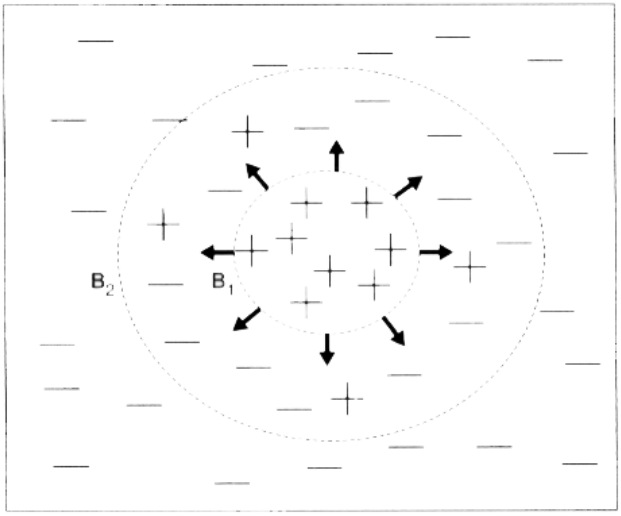
Таблица 13.3. Матриците на несъответствия за два класификационни модела

Обърнете внимание, че макар моделът M_2 е подобрил броя на своите действително положителни и фалшиво положителни предсказвания, той все пак остава по-лош от M_1, тъй като това подобряване става за сметка на увеличаване на по-скъпи фалшиво отрицателни грешки. При използване на стандартната мярка за класификационната точност моделът M_2 би бил по-предпочитан пред модела M_1.

Една класификационна техника с отчитане на цената използва ценовата матрица при построяване на модела като създава модел, имащ най-малката цена. Например, ако най-скъпи са фалшиво отрицателните грешки, то обучаващият алгоритъм ще се опита да намали тези грешки чрез разширяване на границата за взимане на решения в посока към отрицателния клас, като е показано на Фиг. 13.4. По този начин създаденият модел ще може да покрива повече положителни примери, макар че това ще стане за сметка на генериране на допълнителни фалшиви тревоги.

Има различни начини да включим информацията за цената в класификационните алгоритми. Например, в контекста на класификационни дървета ценовата информация може да се използва:

1. За избор на най-добрия атрибут, използван за разделяне на данни
2. За определяне, дали едно под-дърво трябва да бъде изрязано
3. За манипулиране на теглата на обучаващите примери с цел обучаващият алгоритъм да създава класификационното дърво, имащо най-малката цена
4. За модифициране на правилата за избор на класа във всяко листо



Фиг. 13.4. Модификация на границата за вземане на решения (от B_1 към B_2) за намаляване на фалшиво отрицателните грешки на класификатора.

За да илюстрираме последния подход, да обозначим чрез $p(i|t)$ пропорцията на обучаващите примери от класа i, които попадат в листото t. В случая на задача за двоичната класификация, едно типично правило за избор на класа назначава положителния клас в листо t при следните условия:

$$p(+|t) > p(-|t)$$
$$\Rightarrow \quad p(+|t) > (1 - p(+|t))$$
$$\Rightarrow \quad 2p(+|t) > 1$$
$$\Rightarrow \quad p(+|t) > 0.5$$

С други думи се назначава класът, който има най-много примери в листото.

Един алгоритъм, отчитащ цената, ще назначава клас i във възел t по такъв начин, че да минимизира следния израз:

$$C(i \mid t) = \sum_j p(j \mid t) C(j,t)$$

В случая, когато *C(+, +) = C(-, -) = 0*, листото *t* ще назначава положителния клас само, ако:

$$p(+|t)C(+, -) > p(-|t)C(-, +)$$

$$\Rightarrow \quad p(+|t)C(+, -) > (1 - p(+|t))C(-, +)$$

$$\Rightarrow \quad p(+ \mid t) > \frac{C(-,+)}{C(-,+) + C(+,-)}$$

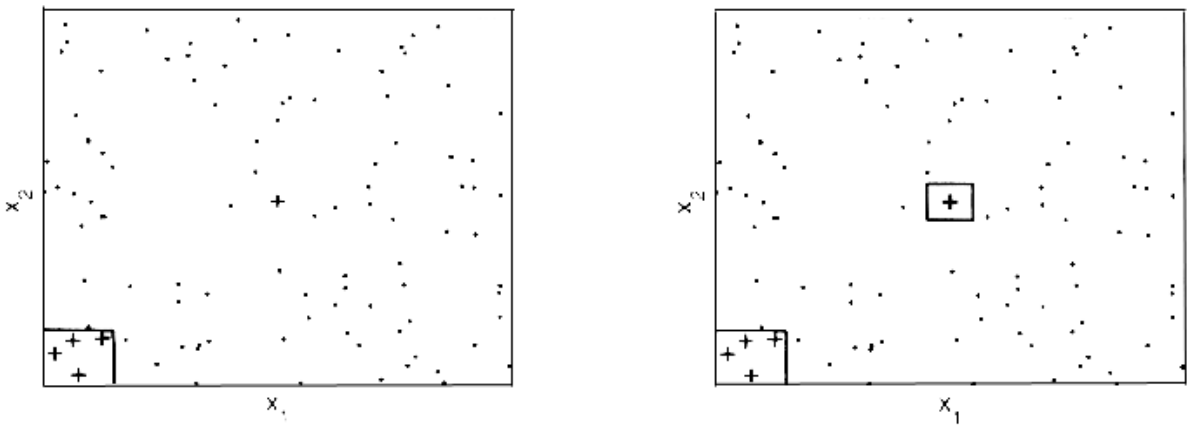
Този израз предлага модифициране на прага за вземане на решения от 0.5 към *C(-, +)/(C(-, +) + C(+, -))*, за да получим един чувствителен към цената класификатор. Ако *C(-, +) < C(+, -)* то прагът ще бъде по-малък от 0.5. Този резултат има смисъл, тъй като цената за направяне на една фалшиво отрицателна грешка е по-голяма от тази, за генериране на една фалшива тревога. Намаляването на прага ще разшири границата за вземане на решения в посока към отрицателния клас, както това е показано на Фиг. 13.4.

13.3.3. Подходи, базирани на извадки

Използването на извадки е един друг широко прилаган подход за справяне с проблема с небалансирани класове. Идеята е да се модифицира разпределението на примери по такъв начин, че рядко срещаният клас да бъде добре представен сред обучаващите примери. Някои от използваните техники са *намаляване на извадката (undersampling)*, *увеличаване на извадката (oversampling)*, както и един хибрид на тези двата подхода. За да илюстрираме тези техники, нека да разгледаме едно множество данни, съдържащо 100 положителни и 1000 отрицателни примера.

В случая на намаляването на извадката, в качество на обучаващото множество се избира една случайна извадка, съдържаща само 100 отрицателни примера и всички положителни примери. Един потенциален проблем с този подход е, че някои от полезни отрицателни примера могат да не попаднат в обучаващото множество, което може да доведе до един не толкова оптимален модел. Потенциалният подход за решаване на този проблем се състои в повтаряне на процедурата по намаляване на извадката няколко пъти и конструиране, съответно, на няколко различни класификатора, които след това могат да бъдат комбинирани в един ансамбъл. Другият възможен подход е да се прави *фокусирано* намаляване на извадката, т.е. процедурата за избор на отрицателните примери ще прави *информирания избор*, елиминирайки, например, тези отрицателни примери, които лежат далеч от границата за вземане на решения.

Процедурата по увеличаване на извадка повтаря *положителните* примери толкова пъти, докато броят на положителните примери в обучаващото множество не стане еднакъв с този на отрицателните примери. Фигура 13.5 илюстрира ефекта от увеличаване на извадката върху създаване на границата за вземане на решения при използване на класификатори от типа на класификационни дървета.



a) Без увеличаване на извадката *b) С увеличаване на извадката*

Фиг. 13.5.Илюстрация на ефекта на увеличаване на извадката на рядко срещания клас

Без увеличаване на извадката само положителните примери, разположени в долния ъгъл на картинката от Фиг. 13.5 а) се класифицират правилно. Положителният пример, разположен в средата на картинката, е класифициран като отрицателен, тъй като няма достатъчно примери за създаване на нова граница за вземане на решения, разделяща положителни от отрицателни примери. Увеличаването на извадката дава допълнителни примери, необходими за създаване на границата около този положителен пример, разделяща класове, както това е показано на Фиг. 13.5 b).

Обаче в случая на зашумени данни увеличаването на извадката може да доведе до преспецификацията на модела, тъй като някои зашумени примери могат да бъдат повторени няколко пъти. По принцип увеличаването на извадката не добавя никаква нова информация към обучаващите данни. Повтарянето на положителните примери само предпазва обучаващия алгоритъм от отрязване на някои части на модела, описващи области, съдържащи прекалено малко обучаващи примери (т.е. малки дизюнкти). Допълнителните положителни примери също така водят до увеличаване на времето за изчисления, необходимо за построяване на модела.

Хибридният подход използва една комбинация от намаляване на броя на примери от най-често срещания клас и увеличаване на примери от най-рядко срещания клас, за да се получи еднакво разпределение на класове. Намаляването може да бъде извършено по случайния или фокусирания начин. Увеличаването може да бъде направено или чрез дублиране на вече съществуващи положителни примери, или чрез създаване на нови положителни примери, намиращи се в околността на вече съществуващи положителни примери. При последния подход отначало трябва да бъдат определени k най-близки съседи на всеки от положителните примери. Един нов положителен пример се генерира в някоя случайна точка, намираща се на линията, съединяваща положителния пример с един от неговите k най-близки съседи. Този процес се повтаря, докато не бъде постигнат желателния брой положителните примери. В отличие от подхода за дублиране на данни, новите примери позволяват да разширят границата за вземане на решения навън от положителния клас, подобно на подхода, показан на Фиг. 13.4. Обаче и този подход може да се окаже много чувствителен към преспецифицирането на модела.

13.4. Откриване на аномалии, базирани на разстояние

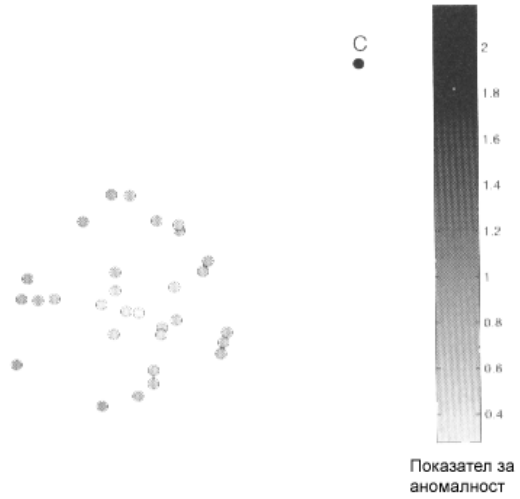
Подходите за откриване на аномалии, базирани на разстояние, не се нуждаят от информация за класа на обекти, както и не построяват никакви модели за описание на нормални (неаномални) данни. Те използват една простичка идея, че един обект е аномален, ако той се намира далеч от повечето други обекти. Единствено от което се нуждае този подход, е дефиниране на някоя марка за близост между обектите.

Един от най-прости начини да се измери, дали един обект се намира далеч от повечето други обекти, е да се използва разстоянието до k-ти неговия най-близък съсед. Този подход определя и дефиницията за крайност, представена по-долу:

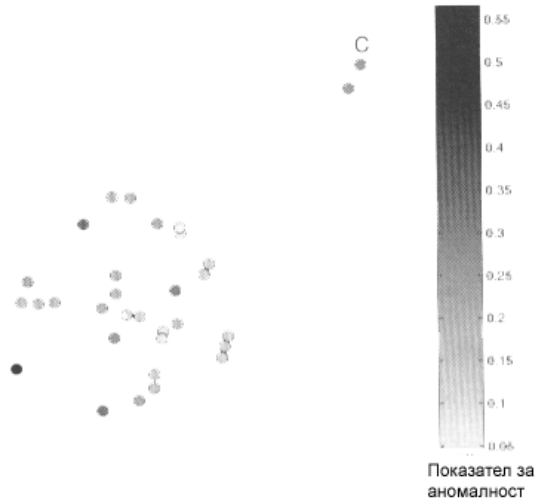
> *Показателят за аномалност на един обект се определя от разстояние до неговия k-ти най-близък съсед.*

Най-малката стойност на такъв показател е 0, докато на-голямата стойност е максимално възможна стойност на функцията на разстояние – обикновено е безкрайност.

На Фиг. 13.6 е показано множеството от двумерни точки. Степента на сивота на всяка от точки указва на стойността на нейния показател за аномалност при $k = 5$. Обърнете внимание, че точка *C* правилно е получила високата стойност на този показател.



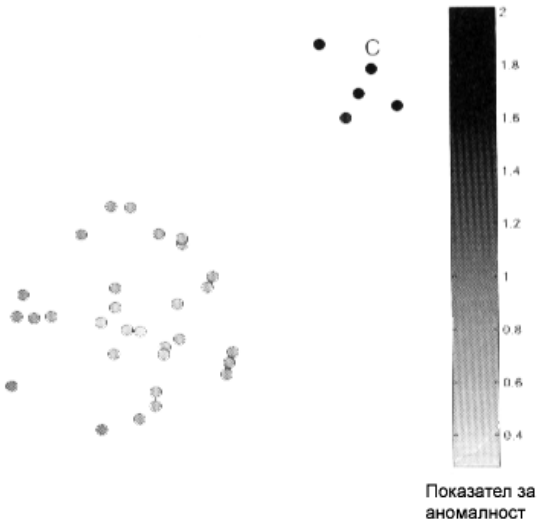
Фиг. 13.6. Показател за анормалност базиран на разстояние до 5-ти най-близък съсед



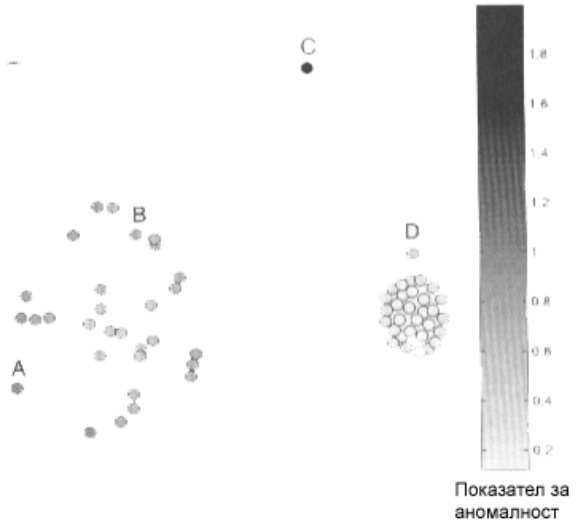
Фиг. 13.7. Показател за анормалност базиран на разстояние до 1-я най-близък съсед. Близките аномалии имат малък показател за аномалност

Показателят за аномалност може да бъде много чувствителен към стойността на *k*. Ако *k* е малко, например 1, то едно малко количество от близко лежащи аномални точки могат да получат малка стойност на този показател. Например Фиг. 13.7 показва множеството от двумерни точки, в което съществува една точка, разположена близо до *C*. Степента на сивото отразява величината на показателя, изчислена при $k = 1$. Обърнете внимание, че *C* и съседната до нея точка имат малки стойности на показателя. Ако *k* е прекалено голямо, то е възможно, че всички обекти в един клъстер, който има по-малко обекти от колко стойността на *k*, се превърнат в крайности.

Например Фиг. 13.8 показва множеството от двумерни точки, което има един естествен клъстер от 5 точки в допълнение към един по-голям клъстер от 30 точки. За $k = 5$ показателят за аномалност на всички точки от малкия клъстер е много висок. За да стане предложената схема по-устойчива към избора на k, предложеното по-горе определение за аномалия трябва да бъде модифицирано, като се използва *средното разстояние* до първите k най-близки съседи.



Фиг. 13.8. Показател за анормалност базиран на разстояние до 5-ти най-близък съсед. Един малък клъстер се превърна в аномалия.



Фиг. 13.9. Показател за анормалност базиран на разстояние до 5-я най-близък съсед. Клъстерите са с различна плътност.

13.4.1. Силни и слаби страни

Описаната по-горе схема за откриване на аномалии, базирана на разстояние, е много проста. Обаче, обикновено подходи, базирани на разстояние, изискват време за изчисление от порядъка на $O(n^2)$, където n е броя на обекти. За големи множества от данни това може да бъде прекалено скъпо, макар че за подобряване на поведението в случая на данни с малка размерност могат да бъдат използвани различни специализирани алгоритми. Освен това, подходът е много чувствителен към избора на параметри. И на края, той не може да се използва за данни, съдържащи области с много различни по величина стойности на плътност, тъй като използва един единствен глобален праг не отчитащ тези промени в плътността.

За илюстрация на този недостатък, да разгледаме множеството от двумерни точки, изобразено на Фиг. 13.9. Множеството се състои от един доста рехав клъстер, един плътен клъстер и две точки – C и D, намиращи се сравнително далече от двата клъстера. Назначавайки на точки стойността на показателя за анормалност, изчислена съгласно предложеното по-горе определение при $k = 5$, точката C ще бъде правилно идентифицирана като анормална, докато точката D ще има малка стойност на този показател. Фактически стойността на показателя за анормалност на точката D ще е значително по-малка от стойността на същия показател за множество от точки, формиращи рехавия клъстер.

13.5. Откриване на аномалии, базирани на плътността

От гледната точка на плътността, аномалните са обекти, намиращи се в области с малка плътност:

> *Показателят за аномалност на един обект е обратен на плътността около обекта.*

Откриването на аномалии, базирани на плътността, е тясно свързано с откриването на аномалии, базирани на разстояние, тъй като плътността е обикновено се дефинира в термините на разстояние. Един общоприет подход е определянето на плътността като стойност, обратна на средното разстояние до k най-близки съседи. Ако това разстояние е малко, то плътността е висока и обратно. Този факт е отразен в следното определение за плътността, като обратното разстояние:

$$Density(\mathbf{x},k)=\left(\frac{\sum_{y\in N(x,k)} dist(\mathbf{x},\mathbf{y})}{|N(\mathbf{x},k)|}\right)^{-1}$$

където $N(\mathbf{x}, k)$ е множество от k най-близки съседи на $\mathbf{x}$, $|N(\mathbf{x}, k)|$ е размер на това множество, а $\mathbf{y}$ е един от най-близките съседи.

Другата дефиниция на плътност е тази, използвана от клъстерния алгоритъм DBSCAN:

> *Плътността около един обект е равна на броя на обекти, намиращи се не по-далеч от разстоянието d от него.*

Параметърът d трябва да бъде избиран много внимателно. Ако d е избрано прекалено малко, то много от нормалните точки могат да имат малка плътност и, следователно, да имат висок показател за анормалност. Ако d е избрано прекалено голямо, то много от аномалните точки ще имат плътности (и показатели за анормалност) сходни с тези на нормалните точки.

Откриването на аномалии чрез използване на всяка от посочените по-горе дефиниции има същите характеристики и ограничения както и тези, отнасящи се за схемите за откриване на аномалии, базирани на разстояние, дискутирани в предишния раздел. В частност, те не могат да идентифицират аномалиите правилно, когато данните съдържат области с различни плътности (виж Фиг. 13.9). За правилната идентификация на аномалии в подобни множества от данни е необходимо да имаме понятието за плътност, свързано с околността на обекта. Например, точката D от Фиг. 13.9 има една по-висока стойност на абсолютната плътност съгласно и двете определения за плътност от колко точката A, но нейната плътност е по-малка в сравнение със същия показател на нейните най-близки съседи.

Има много различни начини да дефинираме относителната плътност на един обект. Един от методите е този, който използва клъстерния алгоритъм, базиран на SNN-плътност, дискутиран подробно в Лекция 7. Другият метод за изчисляване на относителната плътност е във вид на отношението между плътността на някоя точка $\mathbf{x}$ и средната плътност на нейните най-близки съседи $\mathbf{y}$:

$$Average\ relative\ density(\mathbf{x},k) = \frac{Density(\mathbf{x},k)}{\sum_{y \in N(x,k)} Density(\mathbf{y},k) / |N(\mathbf{x},k)|}$$

13.5.1. Откриване на аномалии чрез използване на относителната плътност

В настоящия раздел ще разгледаме една по-опростена версия на техниката, наричана локален фактор за крайности (LOF – Local Outlier Factor). Тя е описана детайлно чрез алгоритъма, представен в Таблица 13.4. На кратко, първоначално алгоритмът изчислява показателя за аномалност за всеки от *k* най-близки съседи чрез изчисляване на плътността на обекта *Density*(**x**, k), базирайки се на неговите най-близки съседи. След това се изчислява средната плътност на съседи на някоя точка, която се използва за изчисляване на относителната плътност на точката, съгласно описаното по-горе уравнение. Тази стойност е индикатор за това, дали **x** се намира в по-гъста или в по-разредена област от околността, отколко нейните съседи, и се взима като стойност на показателя за аномалност на **x**.

Алгоритъм 13.2 – Откриване на аномалии чрез относителната плътност

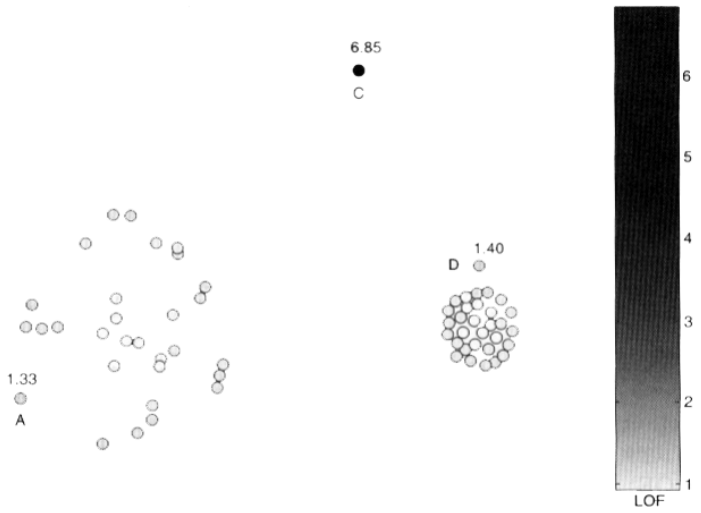
1. Вход: k – броя на най-близки съседи
2. **for all** обекти **x do**
3. Определи *N*(**x**, k) – *k* най-близки съседи на **x**.
4 Определи *Density*(**x**, k) – плътността на **x**, използвайки обекти от *N*(**x**, k)
5. **end for**
6. **for all** обекти **x do**
7. *score*(**x**, k) *:= average relative density*(**x**, k)
8. **end for**

Таблица 13.4. Алгоритъм за откриване на аномалии, базиран на относителната плътност

Пример 13.3. Да разгледаме поведението на описания метод за примерните данни, показани на Фиг. 13.9. Да изберем *k = 10*. Показателят за аномалност на тези точки е показан на Фиг. 13.10. Степента на сивота на всяка точка отразява стойността на нейния показател за аномалност, т.е. точките с по-висок показател са по-тъмни. Точките с най-високия показател са означени с *A, C* и *D*. Това са точки със, съответно, най-високата абсолютна стойност на показателя, с най-високата стойност на показателя относително точки от рехавия клъстер и с най-високия показател относително точки от компактния клъстер.

Откриването на аномалии чрез използване на относителната плътност предоставя една количествена мярка за степента, в която един обект може да се разглежда като аномалия, дори ако данните имат области с различаваща се плътност. Както и подходите, базирани на разстояние, необходимото време за изчисление е, обикновено, от порядъка на $O(n^2)$, макар че за данни с малка размерност то може да бъде намалено до *O(n*log n)* чрез използване на специални структури данни. Освен това изборът на параметрите остава една доста сложна задача, макар че стандартният LOF алгоритъм решава този проблем чрез преглед на множество от стойностите на *k* и последващия

избор на максималните стойности на показателя за аномалност. Обаче, горната и долната граници на тези стойности остават предмет на избора.



Фиг. 13.10. Показателите за аномалност на двумерните точки от Фиг.13.9, изчислени на базата на относителната плътност (LOF)

13.6. Откриване на аномалии чрез техники за клъстеризация

Клъстерният анализ намира групи на силно свързани обекти, докато откриването на аномалии намира обекти, които не са силно свързани с други обекти. По тази причина не буди учудване фактът, че клъстеризацията може да се използва за откриване на аномалии. В този раздел ще разгледаме някои такива техники.

Един от подходите за използване на клъстеризацията за откриване на аномалии се състои в премахване на малки клъстери, когато те са намират далеч от други клъстери. Този подход може да бъде използван с всяка клъстерна техника, но изисква задаването на прагове за минималния размер на клъстера и за разстоянието между един малък клъстер и други клъстери. Често процесът се опростява чрез премахване на всички клъстери, по-малки от зададения минимален размер. Тази схема е силно чувствителна към избрания брой клъстери. Освен това в тази схема е доста трудно да се свърже обектите с техния показател за аномалност. Обърнете внимание, че разглеждането на групи обекти като аномалии представлява разширението на понятието „аномалност" от отделни обекти към групи от обекти, но с нищо не променя същността на понятието.

Един по-систематичен подход се състои в първоначалната клъстеризация на всички обекти, а след това – в оценяването на степента, в която всеки обект принадлежи към съответния клъстер. За клъстеризацията, базирана на прототипи, разстоянието от обекта до центъра на клъстера може да се използва като мярка за принадлежност на обекта към този клъстер. По-общо, за клъстерните техники, които използват някаква обективна мярка, можем да използваме тази мярка за оценяване, до каква степен един обект принадлежи към някакъв клъстер. На пример за алгоритъма k-Means премахването на един обект, разположен далеч от центъра на съответния клъстер, може значително да намали общата квадратична грешка (SSE) на този клъстер. И така,

клъстеризацията създава определен модел за данните, а аномалиите изкривяват този модел. Тази идея е отразена в следното определение за аномалията:

> *Един обект е аномалия, изчислена чрез клъстеризация, ако той не принадлежи в достатъчна степен към никой клъстер.*

На практика, ако използваме някаква клъстерна схема, базирана на определена обективна функция, то това определение е просто един специален случай на определението за аномалност, базирано на модели. Макар че определението е по-естествено за схемите, базирани на прототипи или някаква обективна функция, то може също така включва и подходите към откриване на аномалии, базирани на клъстеризация с използване на плътността или свързаност. В частност за клъстеризацията, базирана на плътността, един обект не принадлежи достатъчно силно към никакъв клъстер, ако неговата плътност е прекалено низка, докато за клъстеризацията, базирана на свързаност, един обект не принадлежи достатъчно силно към никакъв клъстер, ако той не е достатъчно свързан.

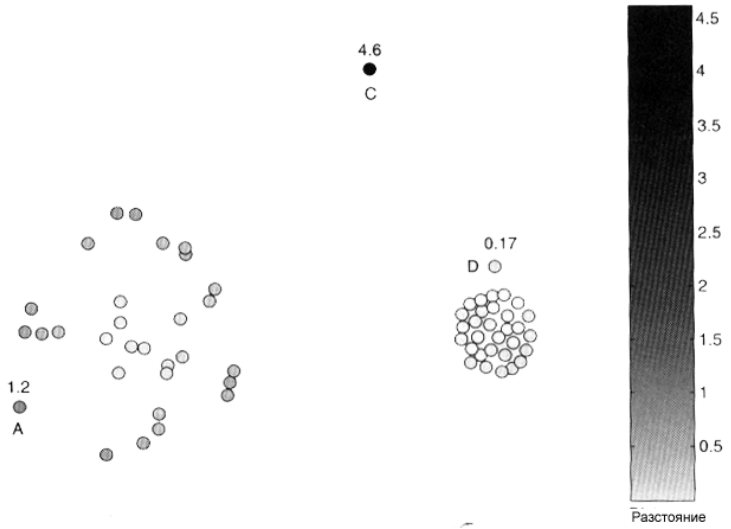
По-долу ще дискутираме няколко въпроса, които трябва да бъдат решени при използване на каквато и да е клъстерната техника за целите на откриване на аномалии. Ще се фокусираме основно на техниките, базирани на прототипи, от типа на k-Means.

13.6.1. Оценка на степента за принадлежност на един обект към клъстера

За клъстери, базирани на прототипи, съществуват няколко начина за оценяване на степента за принадлежността на обекта към клъстера. Един от тях е да се мери разстоянието от обекта до прототипа на клъстера и да се използва това разстояние като показател за аномалност. Обаче, ако клъстерите имат различаващи се плътности, то може да бъде конструирана една друга мярка, измерваща относителното разстояние от обекта до клъстерния прототип, отчитаща разстояния между други обекти от същия клъстер. Една друга възможност, основана на факта, че клъстерите могат да бъдат акуратно моделирани в термините на Гаусови разпределения, е да се използва разстоянието по Махаланобис.

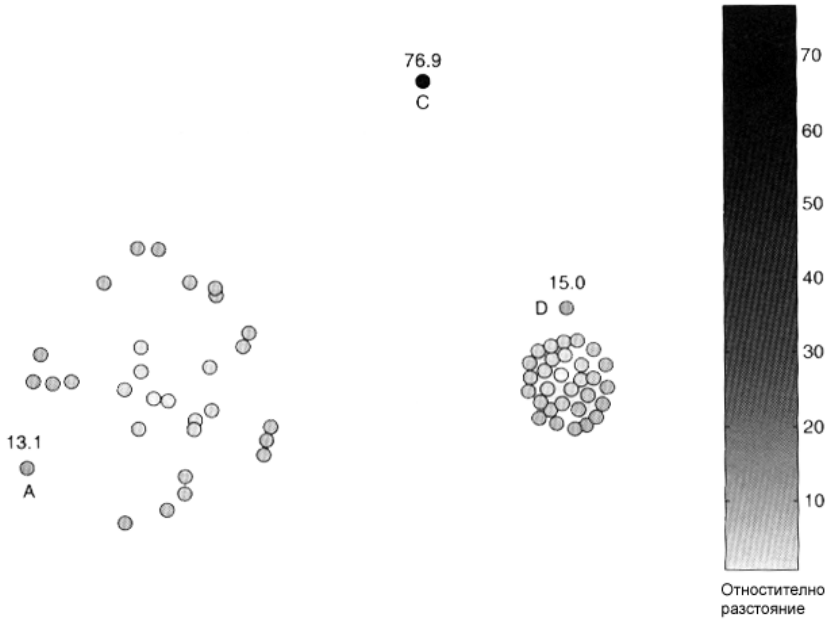
За клъстерните техники, използващи някои от обектните функции, може да назначаваме на обекта показател за аномалност, отразяващ подобрението на обектната функция при елиминирането на обекта. Обаче, оценяването на степента, в която един обект може да се разглежда като аномалия, извършвано на базата на обектната функция, може да се окаже много скъпо от изчислителната гледна точка. По тази причина се предпочитат описаните по-рано подходи, базирани на разстояние.

Пример 13.4. Този пример се базира върху множеството от точки, показани на Фиг. 13.9. За клъстеризацията е използван алгоритмът k-Means, а показателят за аномалност за всяка точка е изчислен по два начина: 1) чрез разстоянието от точка до най-близкия центроид и 2) чрез относителното разстояние от точка до най-близкия центроид, където относителното разстояние е разстоянието от точка до центроида делено върху средното разстояние от всички точки в клъстера до центроида. Последният подход се използва за компенсация на големите разлики в плътността на компактния и рехавия клъстери.



Фиг. 13.11. Разстояния от точки до най-близкия центроид

Получените стойности на показателя за аномалност са показани на Фиг. 13.11 и Фиг. 13.12. Във всеки от случаите са използвани два клъстера. Подходът, базиран само на разстояние, има проблеми с различаващи се плътности на клъстери, например точката *D* не се разглежда като аномалия. При подхода, базиран върху относителното разстояние, всички точки, които по-рано са били идентифицирани като аномалии при използване на LOF алгоритъма (*A, C* и *D*), също са определени като аномалии.



Фиг. 13.12. Относителни разстояния от точки до най-близкия центроид

13.6.2. Влияние на аномалии върху началното клъстеризиране

Ако аномалиите се откриват чрез клъстеризацията, възниква въпросът, дали резултатите са валидни, тъй като самите аномалии влияят върху клъстеризацията. За намиране на отговора на този въпрос можем да използваме следния подход: обектите се

клъстеризират, аномалиите се премахват, а след това останалите обекти отново се подлагат на клъстеризацията. Макар че този подход не гарантира получаване на оптимални резултати, той е много лесен за реализиране. Един по-сложен подход се състои в създаване на една специална група обекти, които в момента не пасват добре на никакъв клъстер. Тази група представлява потенциални аномалии. С напредването на клъстеризационния процес променят се и самите клъстери. Обектите, за които вече стана ясно, че не принадлежат към нито един от клъстери, се добавят към множеството от потенциални аномалии, докато другите обекти от това множество се проверяват, дали те вече не принадлежат достатъчно силно към някой от клъстери, и ако е така, то те се премахват от множеството на потенциални аномалии. Обектите, останали в това множество след края на клъстеризационния процес, се класифицират като аномалии. Отново, няма никаква гаранция за получаване на оптималното решение или дори за това, че този подход ще работи по-добре от по-простия подход, описан по-рано. Например, един клъстер от зашумени точки може да изглежда като истинския клъстер без никакви аномалии. Този проблем е особено сериозен, ако показателят за аномалност се изчислява на базата на относителното разстояние.

13.6.3. Влияние на броя на клъстери

Клъстерните техники от рода на k-Means не избират автоматично броя на клъстери. Това води до проблеми, когато клъстеризацията се използва за определяне на аномалии, тъй като дали един обект ще се разглежда като аномалия или не, може да зависи от избрания брой на клъстери. Например в една група от 10 обекта самите обекти могат да бъдат разположени относително близо един до друг, но не зависимо от това, те могат и да бъдат включени като част в един по-голям клъстер, ако са намерени само няколко големи клъстери. В този случай десетте точки могат да се разглеждат като една аномалия, макар че те могат да формират и един отделен клъстер, ако достатъчно голям брой клъстери са били указани като параметър на клъстеризацията.

Както и при други подобни проблеми, не съществува един прост отговор на този проблем. Една от възможни стратегии е да бъде повторен анализът при различен брой на клъстери. Другият подход се състои в намирането на достатъчно голям брой малки клъстери. Идеята е, че 1) по-малките клъстери имат тенденция да бъдат по-компактни и 2) ако един обект е аномалия дори при голямото количество от малки клъстери, то има голяма вероятност, че той е истински аномален обект. Обаче при този подход групите от аномалии, които формират малки клъстери, могат да останат незабелязани.

13.6.4. Силни и слаби страни

Някои клъстерни техники, такива като k-Means, имат линейна или близо до линейна изчислителна сложност и, по тази причина, откриването на аномалии, базирани на такива алгоритми, е много ефективно. Освен това, дефиницията на един клъстер е често допълваща до тази на една аномалия и, по тази причина, обикновено е възможно намирането на клъстери и аномалии да става по едно и също време. Като слабите страни на този подход трябва да се отбележат, че множеството от аномалии и стойностите на показателя за аномалност на съответните точки в много голяма степен зависят от избрания брой на клъстери и от наличието на истинските аномалии в анализираните данни. Например клъстерите, намерени от един базиран на прототипи

алгоритъм, могат да бъдат изкривени от наличието на аномалии. Качеството на аномалии, намерени от един клъстерен алгоритъм, е силно повлияно от качеството на клъстери, намерени от този алгоритъм. Както сме дискутирали в лекциите, посветени на клъстеризацията, всеки клъстерен алгоритъм е подходящ само за определен тип данни; следователно и при откриването на аномалии чрез клъстеризация, самият клъстерен алгоритъм трябва да бъде избиран много внимателно.