Data Mining: Concepts and Techniques

(3rd ed.)

— Chapter 12 —

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign &
Simon Fraser University

Chapter 12. Outlier Analysis

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

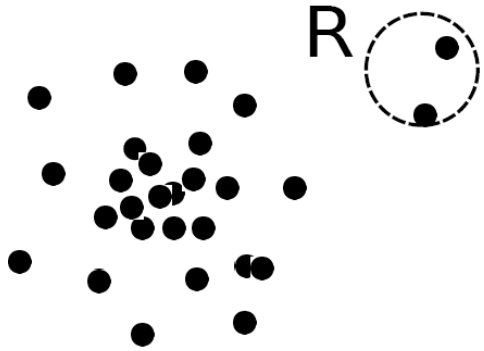
- این جلسه متفاوت از جلسه حضوری است. میخواهیم فصل ۱۲ را تمام کنیم. در مورد outlier detection صحبت می کنیم. outlier به معنی دور افتاده است یا پرت. هدف ما از داده کاوی پیدا کردن داده های پرت است و نه تنها داده های پرت را پیدا می کنیم باید یک استراتژی برخورد با آنها را برای سیستم یا برای خودمان که داده کاوی انجام می دهیم در نظر بگیریم. این فصل از کتاب هان گفته می شود ویرایش سوم فصل ۱۲ هان. کتاب هان را هم دانشگاه تهران ترجمه کرده که میشه فصل ۵ آن. تمرینات فصل ۱۲ را باید حل کنیم.داده ای دور افتاده یا تحلیل دورافتاده.
- چرا داده ای دورافتاده واسه ما در داده کاوی اهمیت دارد با یک مثال شرح داده. فرض کنید ما مسـُول بررسی خرید از کارت اعتباری هستیم. خیلی وقت ها کارت گم می شود یا سایتی حک می شود و از کارت استفاده بی جا می کنند. برای کم کردن این استفاده ها خود بانک پیامک میزند به کسی که از کارتش استفاده شده اما ممکن است پیامک را دریافت نکند و روند استفاده بی جا از کارت ادامه پیدا کند این اتفاق شاید برای فرد یک اتفاق نادر است. وقتی خرید بالایی را از کارت می کنیم که در شهر خودش نیست و جای دور است.

- باید در پیدا کردن این outlier detection ما حواسمان باشد که خطایی رخ ندهد اگر خطا رخ داد هزینه ای که ما تحمیل می شویم زیاد است
- ما انواع outlier و روش های پیدا کردن آن ها را می گوییم.
- قدیمی ترین روشها، روشهای آماری است. کلاسترینگ و کلاسیفیکیشن دو روش عمده برای پیدا کردن outlier هستند.
- تعریف داده پرت :
- داده ای است که با یک مکانیزم متفاوتی از داده های معمولی یا نرمال تولید شده که به آن abnormal می گویند. نرمال می تواند توزیع نرمال باشد منظور از نرمال در اینجا یعنی عادی و عرف و abnormal غیر عادی است. مثل استفاده از کارت هنگام خرید که متفاوت از همیشه باشد . دور افتاده ها با دادهای نویزی متفاوت هستند. نویز، نوفه یا آشوب خطایی است که بر اساس نمونه گیری بر اساس مواردی که نمیتوانیم کنترل کنیم. اما outlier بخشی از داده ما است و نمی توان آن را به عنوان نویز تلقی کرد. نویز تصادفی است. قبل از بحث outlier باید نویز را حذف کنیم. نویز خطای سیستم است که در مدل در نظر می گیریم و مقدار زیادی ندارد .

- اما outlier اینطوری نیست .اگر ما بتوانیم outlier ها را مشخص کنیم که درصد آنها کمتر از داده های نرمال هستند می توانیم به طور غیر مستقیم داده های نرمال را مشخص کنیم.

- برای پیدا کردن outlier ها دو روش داریم ابتدا داده های نرمال را مشخص کنیم و بقیه را برچسب outlier بزنیم یا اول داده های دور افتاده را مشخص کنیم و بقیه را برچسب نرمال بودن بگذاریم. اینکه بخواهیم تصادفی بودن را آزمون کنیم امکان دارد به شرطی که قبل آن بتوانیم مدل را به داده ها برازش کنیم.
- مثلا در سری زمانی اگر یک مدل سری زمانی به داده ها برازش دهیم از داده ها مدل را کم می کنیم مثلا x_tاسم داده های ما باشد چیزی که ما برازش می دهیم x^t است همان کاری که در itsm می کردیم. حال x_tرا منهای x^t بکنیم که Errorرا به ما می دهد. Error یک مشاهده از وایت نویز است که به آن residual می گوییم به وایت نویز Errorمی گوییم و residual برآورد Errorاست. Errorقابل دیدن نیست می توان مقدار ان را تقریبی مشخص کرد.

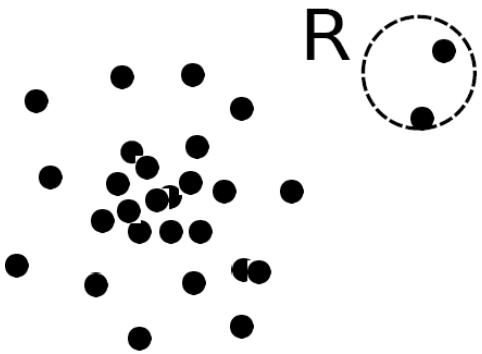
- مدل را که برازش دادیم برای هر مشاهده یک مقدار predicted دارد که پیش بینی ما x^t می شود به x_t منهای x^t ، Residual می گویند. مانده یک مصداقی از خطایمان است می توان خطا را تست کرد که تصادفی است یا خیر. اگر تصادفی بود یعنی مدل را خوب برازش دادیم و اگر در مدل بخشی برای outlierها در نظر نگرفتیم و آنها را حذف نکردیم نتیجه می گیریم که outlier نداشتیم.
- outlier detection با novelty detection متفاوت است و بعضی اوقات با روند جدیدی که پیش می اید متفاوت است مثلا در کارت بانکی یک اتفاقی در کشور می افتد که از کارت بانکی اشان استفاده دیگری هم از هفته آینده می توانند بکنند و این استفاده از کارت را نمیتوانیم به عنوان outlier بگذاریم. در فصل یک این ها را بحث کردیم. از کاربردهای آن کشف تبهکاری در پزشکی برای داده ای دو بعدی است که داده ای که با R نشان داده با بقیه متفاوت است.



- در رگرسیون ما دو جور داده دور افتاده داریم داده هایی که پایین خط رگرسیون که به آن ها نقطه اهرمی میگفتیم(leverage point) و داده های بالای خط رگرسیون (outlier) میگفتیم معمولا این گونه داده ها را نرم افزارها علامت گذاری می کنند و بعد از برازش مدل رگرسیون به ما ارایه می دهند و داده های پرت با مشکوک را مشخص می کنند.

What Are Outliers?

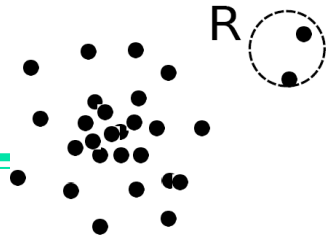
- **Outlier**: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism**
  - Ex.: Unusual credit card purchase, sports: Michael Jordon, Wayne Gretzky, ...
- Outliers are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection
- Outliers are interesting: It violates the mechanism that generates the normal data
- Outlier detection vs. *novelty detection*: early stage, outlier; but later merged into the model
- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis



Types of Outliers (I)

- ما سه جور داد ه پرت داریم ۱- global داده سراسری ۲- contextualداده ضمیمه ای یا شرطی ۳- collective داده های اجماعی.
- داده global مقدارش با بقیه داده ها به طور فاحشی متفاوت است برای پیدا کردن این داده ها باید متری داشته باشیم که بتوان با متر انحراف داده را با داده های نرمال تشخیص داد معمولا واریانس داده های معمول استفاده می شود.
- در داده های ضمیمه ای باید یک ضمینه ای از مساله ای که مطرح شده همراه عدد بدهیم. مثلا درجه حرارت تورنتو ۲۸ است اگر مساله را در زمستان بررسی میکنیم باید ۲۸ درجه زیر صفر باشد نه بالای صفر . و اگر به ما فصل را نمی گفتن پس ۲۸ درجه در تابستان در تورنتو ممکن است. اگر صورت مساله را ندانیم نمیتوانیم تشخیص دهیم که outlier است یا نه . هر سوالی که مطرح میشه باید مانند زمان و مکان را بررسی کنیم. هم تورنتو مهم است هم فصلش.
- Collective outlier دورافتاده های اجماعی داده های تعدادی هستند و تک نیستند و تعدادش شاید به نسبت کل داده ها کم باشد

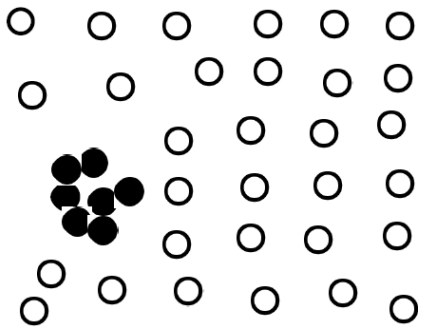
Types of Outliers (I)



Global Outlier

- Three kinds: *global, contextual* and *collective* outliers
- **Global outlier** (or point anomaly)
  - Object is O_g if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation
- **Contextual outlier** (or *conditional outlier*)
  - Object is O_c if it deviates significantly based on a selected context
  - Ex. 80° F in Urbana: outlier? (depending on summer or winter?)
  - Attributes of data objects should be divided into two groups
    - Contextual attributes: defines the context, e.g., time & location
    - Behavioral attributes: characteristics of the object, used in outlier evaluation, e.g., temperature
  - Can be viewed as a generalization of *local outliers*—whose density significantly deviates from its local area
  - Issue: How to define or formulate meaningful context?

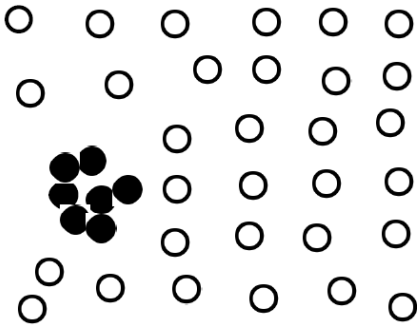
- مثال در شکل تعدادی از داده ها الگوی ان ها با بقیه متفاوت است در این حالت علاوه بر اینکه تک تک داده ها را برایشان outlier بررسی میکنیم باید به طور مجزا انها را به صورت گروهی انها را آزمون کنیم که outlier هستند یا نه. یک dataset ممکن است هر سه جور outlierرا داشته باشد. میتوان یک داده ای داشت که هم global و هم contextual و هم collective outlierباشد



Collective Outlier

Types of Outliers (II)

- **Collective Outliers**
  - A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers
  - Applications: E.g., *intrusion detection*:
    - When a number of computers keep sending denial-of-service packages to each other
  - Detection of collective outliers
    - Consider not only behavior of individual objects, but also that of groups of objects
    - Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects.



Collective Outlier

- A data set may have multiple types of outlier
- One object may belong to more than one type of outlier

Challenges of Outlier Detection

- ما دشواری هایی را داریم برای اینکه بتوانیم outlierها را پیدا کنیم اینکه نرمال بودن را تعریف کنیم کار ساده های نیست و یک حد آستانه ای مشخص کنیم که آن هایی که آن سمت خط هستند نرمال و آن هایی که این سمت هستند غیر نرمال هستند.
- پیدا کردن نرمال و غیر نرمال سخت است . مشکلاتی که ما در پیدا کردن outlierها با آن ها برخورد می کنیم متری است که اندازه گیری میکنیم. مثال: اگر ما متر را توان دو در نظر بگیریم نسبت به داده های پرت و leverage pointها حساس است. اگر یک داده leverage point داشته باشیم خط رگرسیون سمت آن نقطه میرود. ولی اگر absolute error log در نظر بگیریم نسبت به نقاط leverage point که پایین خط رگرسیون می افتد حساس نیست. در آمار داده های پرت را پیدا و حذف میکنم. اما در داده کاوی داده های پرت را پیدا میکنیم و خدمت آن ها میرسیم که چه اتفاقی برای ان ها افتاده. روش شاید در آمار و داده کاوی یکی باشد اما برخورد ما با داده های چرت متفاوت است.

- ما در کارهای مختلف نحوه برخوردمان با داده های پرت متفاوت است. مثلا در داده های کلنیکی اگر کوچکترین تغییر در فشار یا قند خون پیدا شود باید سریع کاری کنیم در صورتی که مقادیر اگر در بازاری رخ دهد برایمان مهم نیست عادی است.flactuation اگر در بازاری رخ دهد برای ما مهم نیست عادی است. قابل فهم بودن آنها واسه ما این هست که باید دلیلی آورد که چرا داده ها outlier شدند. باید دلیل را بعد از پیدا کردن داده پرت گفت. باید درجه ای از درست نمایی یا احتمال پرت بودن را هم بیان کنیم.

Challenges of Outlier Detection

- Modeling normal objects and outliers properly
  - Hard to enumerate all possible normal behaviors in an application
  - The border between normal and outlier objects is often a gray area
- Application-specific outlier detection
  - Choice of distance measure among objects and the model of relationship among objects are often application-dependent
  - E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations
- Handling noise in outlier detection
  - Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection
- Understandability
  - Understand why these are outliers: Justification of the detection
  - Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism

Outlier detection method

- روش های تشخیص و پیدا کردن outlierها
- معمولا تمام روش های پیدا کردن outlierرا میتوان به دو کلاس عمده تقسیم کرد . برچسب های outlierرا میتوان بر اساس یک روش با نظر و ناراهنماییده پیدا کنیم. یا بر اساس فرض هایی که در داده های نرمال به ما دادند میتوان با بی ناظر اینها را پیدا کرد روش های آماری و کلاسترینگ در این بخش هستند.
- در روش های راهنمایی به جای کلاسترینگ سراغ کلاسیفیکیشن میرویم چون از اول برچسب نمونه ای از داده های outlier یا داده های نرمال را داریم و مساله را با ناظر میگذاریم در روش های با ناظر ۲۰ ٪ داده ها برای تست و ۸۰ ٪ برای داده های آموزشی و با استفاده از داده های آموزشی مدل برازش میدهیم و با استفاده از داده های تست می بینیم که مدل ما مناسب است یا نه. می خواهیم ببینیم که داده های ما داده دور افتاده دارند یا نه.

- داده ها را به دو کلاس تقسیم می کنیم داده های دورافتاده و نرمال. و یکسری از ان ها که برای train است برچسب ان ها را داریم. از حالا به بعد هر داده ایی دادند با توجه به داده های تست و آموزش میتوان برچسب مناسبی به عنوان داده های پرت و نرمال به آن ها زد.. این داده هایی که درآوردیم برای داده های جدید از ان ها استفاده می کنیم و برچسب میدهیم.
- مشکل ما در این مساله که با کلاسترینگ و کلاسیفیکیشن متفاوت است این است که تعداد outlierها خیلی کمتر از تعداد داده های نرمال است . مثل ژنتیک. مثل بیماری ها. باید تعداد داده ها در هر کلاس یک تعداد معقولی باشد. در اینجا تعداد outlierها نسبت به داده های اصلی کمتر است و برای مدل کردن مساله ما را دچار مشکل میکند.
- باید تعداد آنها قابل قبول باشد .

- هرچقدر تعداد outlierها ی بیشتری داشته باشیم بهتر میتوانیم مدل را برازش دهیم.
- در روش بی ناظر مساله کلاسیفیکیشن به کلاسترینگ تبدیل میشود یعنی ما داده آموزشی نداریم از قبل برچسب نداریم و همان جا باید داده ها را به چند کلاس که شاید یکی از کلاس ها غیر نرمال و بقیه نرمال هستند تقسیم می شود. انتظار داریم یک گروه داشته باشیم یا خوشه که با بقیه خوشه ها اختلاف داشته باشد . محدودیت روش این که collective outlier را به راحتی نمیتوان پیدا کرد چون ما بیس را برای پیدا کردن داده هایی که از گروه ها دور هستند گذاشتیم

- مثال: مثل پیدا کردن یک ویروس که کار راحتی نیست. روش بی ناظر ممکن است high false positive rateداشته باشند. یعنی اینکه outlierاست و ما بگوییم که outlier نباشد. روش های با ناظر کاراتر هستند. تعداد روش های بی ناظر که داریم مثل کلاسترینگ های مختلف میتوان آنها را برای پیدا کردن اینجور داده ها وفق داد. ما دو تا مساله داریم : تشخیص نویز از outlier که هزینه بر است. و زمانی که outlierما تفاوت چندانی با داده های نرمال نداشته باشند.

Chapter 12. Outlier Analysis

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

Outlier Detection I: Supervised Methods

- Two ways to categorize outlier detection methods:
  - Based on <u>whether user-*labeled* examples of outliers can be obtained</u>:
    - Supervised, semi-supervised vs. unsupervised methods
  - Based on <u>*assumptions about normal data and outliers*</u>:
    - Statistical, proximity-based, and clustering-based methods
- **Outlier Detection I: Supervised Methods**
  - Modeling outlier detection as a classification problem
    - Samples examined by domain experts used for training & testing
  - Methods for Learning a classifier for outlier detection effectively:
    - Model normal objects & report those not matching the model as outliers, or
    - Model outliers and treat those not matching the model as normal
  - Challenges
    - Imbalanced classes, i.e., outliers are rare: Boost the outlier class and make up some artificial outliers
    - Catch as many outliers as possible, i.e., recall is more important than accuracy (i.e., not mislabeling normal objects as outliers)

Outlier Detection II: Unsupervised Methods

- Assume the normal objects are somewhat ``clustered'' into multiple groups, each having some distinct features
- An outlier is expected to be far away from any groups of normal objects
- Weakness: Cannot detect collective outlier effectively
  - Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area
- Ex. In some intrusion or virus detection, normal activities are diverse
  - Unsupervised methods may have a high false positive rate but still miss many real outliers.
  - Supervised methods can be more effective, e.g., identify attacking some key resources
- Many clustering methods can be adapted for unsupervised methods
  - Find clusters, then outliers: not belonging to any cluster
  - Problem 1: Hard to distinguish noise from outliers
  - Problem 2: Costly since first clustering: but far less outliers than normal objects
    - Newer methods: tackle outliers directly

- Semi supervised

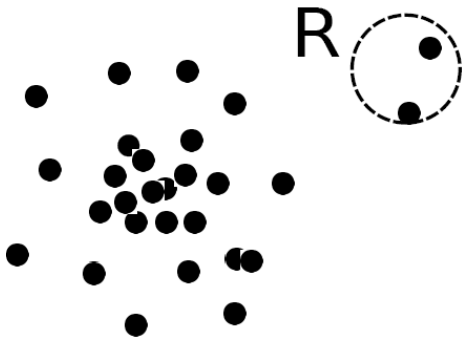
- روشی که نه باناظر است نه بی ناظر . این روشها معقول تر هستند نه خوشبینانه مثل کلاسیفیکیشن نه مثل کلاسترینگ که فقط یکسری اطلاعات جزیی به ما میدهند که با آن کار میکنیم. در اینجا یکسری برچسب را به ما میدهند. یا اینکه میگویند در داد ه ها چند برچسب دارند یا داده ها به پنج خوشه تقسیم میشود. اگر خوشه ششم بود معلوم است که یک اتفاقی افتاده . داده های که در این ۵ کلاسی که داشتیم قرار گرفت میشه outlier.

Outlier Detection III: Semi-Supervised Methods

- Situation: In many applications, the number of labeled data is often small: Labels could be on outliers only, normal objects only, or both
- Semi-supervised outlier detection: Regarded as applications of semi-supervised learning
- If some labeled normal objects are available
  - Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
  - Those not fitting the model of normal objects are detected as outliers
- If only some labeled outliers are available, a small number of labeled outliers many not cover the possible outliers well
  - To improve the quality of outlier detection, one can get help from models for normal objects learned from unsupervised methods

Outlier Detection (1): Statistical Methods

- Statistical methods (also known as model-based methods) assume that the normal data follow some statistical model (a stochastic model)
  - The data not following the model are outliers.
- Example (right figure): First use Gaussian distribution to model the normal data
  - For each object y in region R, estimate $g_D(y)$, the probability of y fits the Gaussian distribution
  - If $g_D(y)$ is very low, y is unlikely generated by the Gaussian model, thus an outlier



- Effectiveness of statistical methods: highly depends on whether the assumption of statistical model holds in the real data
- There are rich alternatives to use various statistical models
  - E.g., parametric vs. non-parametric

Outlier Detection (1): Statistical Methods

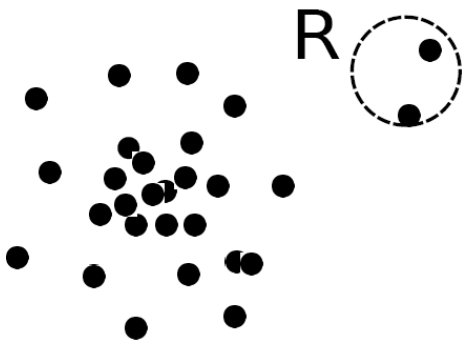
- معرفی چندین روش آماری
- داده هایی که در آمار کار می کنیم خیلی از انها از توزیع نرمال پیروی می کنند مثل نمره دانشجو، طول قد. قضیه حد مرکزی گارانتی می کند که میانگین داده ها توزیع آن نرمال است و وابسته به توزیع نیست. ما مساله توزیع نرمال را به این مساله گسترش می دهیم که اگر در داده ها از توزیع نرمال پیروی نکردند آن ها outlier هستند. می توان مساله را کمی کرد داده هایی که از توزیع نرمال پیروی کنند داده هایی هستند که بیش از ۹۹٪ بین -۳ و ۳ قرار می گیرند اگر بین این ها نبود داده ها outlierهستند. روش های ساده تر مثلا box plotکه رسم می کردیم حالت یک متغیره q1 q2 q3 را پیدا می کردیم یک جعبه می کشیدیم که q3 بالای جعبه q1پایین جعبه q2میانه.

- Q2 q3چارک اول و دوم است. حالا، q3منهای q2 را دامنه میان چارکی بزاریم و دوبرابر میان چارکی یک خط بالا و پایین بکشیم هر داده های که بیشتر از دو برابر میان چارکی بود به آن داده پرت می گوییم که یک روش نا پارامتری است و توزیع آن مشخص نیست و نرمال نیست. توزیع گاوسی نقش مهمی دارد چون میانگین داده ها توزیع نرمال هستند می توان از این روش استفاده کرد . روشهای آماری وقتی کارایی دارند که فرض های اولیه در انها رعایت شود مثل نرمال و استقلال و هم توزیع بودن .می توان روی هر بعد جداگانه این کار را انجام داد.

- مثلا اگر یک داده در حالت دو بعدی داشته باشیم و آن را روی محور ترسیم کنیم حالت پرت بودن آن از بین میرود. اگر داده در حالت یک متغیره پرت بود در دو متغیره یک داده پرت است. وقتی کناری مارجینال به دست میاوریم و ترسیم می کنیم پرت بودن داده از بین میرود. اگر دایره در شکل کمی به نقاط نزدیکتر بود هنگام ترسیم داده پرت بودن آنها را تشخیص نمیدادیم و در محور y قاطی داده ها میشد. در ابعاد بالاتر outlierها خودشان را نشان میدهند. داده ها تنک تر و اسپارس تر میشوند و بد رفتاری آنها مشخص می شود.
- روش ناپارامتری یعنی توزیع جامع مشخص نیست که نرمال باشد یا نه.

Outlier Detection (2): Proximity-Based Methods

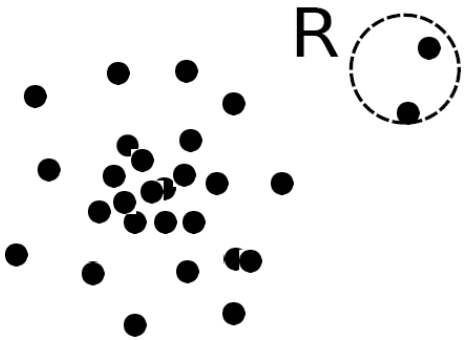
- An object is an outlier if the nearest neighbors of the object are far away, i.e., the **proximity** of the object is **significantly deviates** from the proximity of most of the other objects in the same data set
- Example (right figure): Model the proximity of an object using its 3 nearest neighbors
  - Objects in region R are substantially different from other objects in the data set.
  - Thus the objects in R are outliers



- The effectiveness of proximity-based methods highly relies on the proximity measure.
- In some applications, proximity or distance measures cannot be obtained easily.
- Often have a difficulty in finding a group of outliers which stay close to each other
- Two major types of proximity-based outlier detection
  - Distance-based vs. density-based

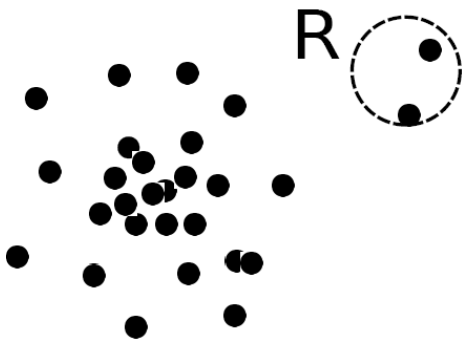
Outlier Detection (2): Proximity-Based Methods

- روش دوم روش مجاورت
- در مثال نزدیکترین همسایگی را مشخص کرده که دو تا نقطه در یک همسایگی و به عنوان outlier قرار گرفتند . در بعضی مسایل نمی توان مجاورت را محاسبه کرد. درمواقعی که داده های outlier به هم نزدیک هستند روشی نمی تواند جواب دهد. ما روش های distanced based و dencsity based هستند که مطرح شده.
- Distanced baed روشهایی هستند که بر اساس کلاسترینگ مطرح می شوند از کلاسترینگ سلسله مراتبی تا روش های دیگر.



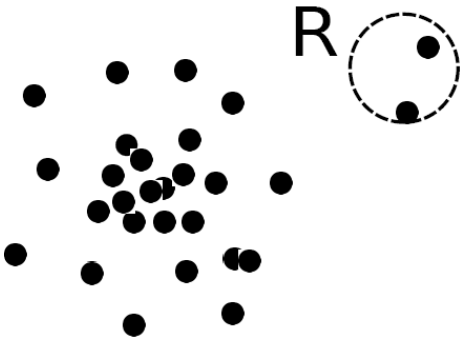
Outlier Detection (3): Clustering-Based Methods

- می خواهیم داده ها را به دو بخش کلاستر کنیم یک بخش خیلی گنده و بخش کوچک. دو خصوصیت دارد. تعداد آنها کم است یا تعداد آن ها خوب است و دو نقطه ای که در R هستند فاصله آن ها کم است. پس یک مشکلی وجود دارد که داده پرت ما اگر یکی جهت دیگری بود انوقت به سه کلاس می رسیم. پس با دو کلاس به نتیجه رسیدن کار جالبی نسیت. روش های کلاسترینگ را برای پیدا کردن outlierها adapt کردند. بهترین روش کلاسترینگ ما سلسله مراتبی است که گران است.
- روشهای آماری را از روی کتاب بخونیم. فصل ۱۲ چاپ سوم کتاب هان .



Outlier Detection (3): Clustering-Based Methods

- Normal data belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters
- Example (right figure): two clusters
  - All points not in R form a large cluster
  - The two points in R form a tiny cluster, thus are outliers



- Since there are many clustering methods, there are many clustering-based outlier detection methods as well
- Clustering is expensive: straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets

Chapter 12. Outlier Analysis

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

Statistical Approaches

- Statistical approaches assume that the objects in a data set are generated by a stochastic process (a generative model)
- Idea: learn a generative model fitting the given data set, and then identify the objects in low probability regions of the model as outliers
- Methods are divided into two categories: *parametric* vs. *non-parametric*
- Parametric method
  - Assumes that the normal data is generated by a parametric distribution with parameter θ
  - The probability density function of the parametric distribution $f(x, \theta)$ gives the probability that object x is generated by the distribution
  - The smaller this value, the more likely x is an outlier
- Non-parametric method
  - Not assume an a-priori statistical model and determine the model from the input data
  - Not completely parameter free but consider the number and nature of the parameters are flexible and not fixed in advance
  - Examples: histogram and kernel density estimation

Parametric Methods I: Detection Univariate Outliers Based on Normal Distribution

- Univariate data: A data set involving only one attribute or variable
- Often assume that data are generated from a normal distribution, learn the parameters from the input data, and identify the points with low probability as outliers
- Ex: Avg. temp.: {24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4}
  - Use the maximum likelihood method to estimate μ and σ

$$\ln \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{n} \ln f(x_i|(\mu, \sigma^2)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - Taking derivatives with respect to μ and σ^2, we derive the following maximum likelihood estimates

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  - For the above data with n = 10, we have $\hat{\mu} = 28.61 \quad \hat{\sigma} = \sqrt{2.29} = 1.51$
  - Then (24 – 28.61) /1.51 = – 3.04 < –3, 24 is an outlier since

$$\mu \pm 3\sigma \text{ region contains } 99.7\% \text{ data}$$

Parametric Methods I: The Grubb's Test

- Univariate outlier detection: The Grubb's test (maximum normed residual test) — another statistical method under normal distribution
  - For each object x in a data set, compute its z-score: x is an outlier if

$$z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

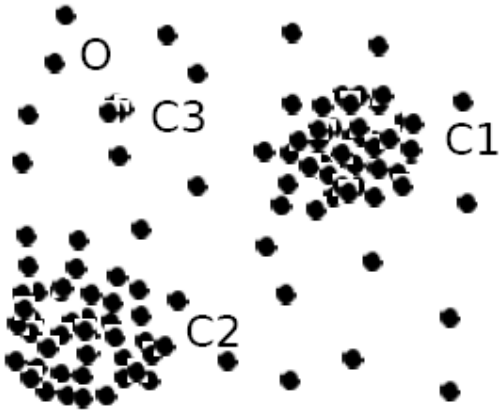
where $t^2_{\alpha/(2N),N-2}$ is the value taken by a t-distribution at a significance level of α/(2N), and N is the # of objects in the data set

Parametric Methods II: Detection of Multivariate Outliers

- Multivariate data: A data set involving two or more attributes or variables
- Transform the multivariate outlier detection task into a univariate outlier detection problem
- Method 1. Compute Mahalaobis distance
  - Let $\bar{o}$ be the mean vector for a multivariate data set. Mahalaobis distance for an object o to $\bar{o}$ is $MDist(o, \bar{o}) = (o - \bar{o})^T S^{-1}(o - \bar{o})$ where S is the covariance matrix
  - Use the Grubb's test on this measure to detect outliers
- Method 2. Use χ^2 –statistic: $\chi^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$
  - where E_i is the mean of the i-dimension among all objects, and n is the dimensionality
  - If χ^2 –statistic is large, then object o_i is an outlier

Parametric Methods III: Using Mixture of Parametric Distributions

- Assuming data generated by a normal distribution could be sometimes overly simplified
- Example (right figure): The objects between the two clusters cannot be captured as outliers since they are close to the estimated mean



- To overcome this problem, assume the normal data is generated by two normal distributions. For any object o in the data set, the probability that o is generated by the mixture of the two distributions is given by

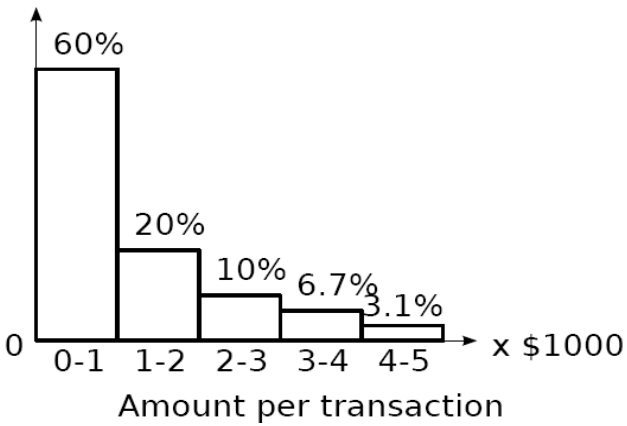
$$Pr(o|\Theta_1, \Theta_2) = f_{\Theta_1}(o) + f_{\Theta_2}(o)$$

where $f_{\theta 1}$ and $f_{\theta 2}$ are the probability density functions of θ_1 and θ_2

- Then use EM algorithm to learn the parameters μ_1, σ_1, μ_2, σ_2 from data
- An object o is an outlier if it does not belong to any cluster

Non-Parametric Methods: Detection Using Histogram

- The model of normal data is learned from the input data without any *a priori* structure.
- Often makes fewer assumptions about the data, and thus can be applicable in more scenarios
- Outlier detection using histogram:



  - Figure shows the histogram of purchase amounts in transactions
  - A transaction in the amount of $7,500 is an outlier, since only 0.2% transactions have an amount higher than $5,000
- Problem: Hard to choose an appropriate bin size for histogram
  - Too small bin size → normal objects in empty/rare bins, false positive
  - Too big bin size → outliers in some frequent bins, false negative
- Solution: Adopt kernel density estimation to estimate the probability density distribution of the data. If the estimated density function is high, the object is likely normal. Otherwise, it is likely an outlier.

Chapter 12. Outlier Analysis

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

Proximity-Based Approaches: Distance-Based vs. Density-Based Outlier Detection

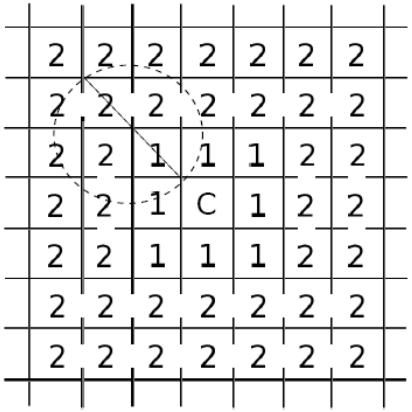
- Intuition: Objects that are far away from the others are outliers
- Assumption of proximity-based approach: The proximity of an outlier deviates significantly from that of most of the others in the data set
- Two types of proximity-based outlier detection methods
  - Distance-based outlier detection: An object o is an outlier if its neighborhood does not have enough other points
  - Density-based outlier detection: An object o is an outlier if its density is relatively much lower than that of its neighbors

Distance-Based Outlier Detection

- For each object o, examine the # of other objects in the *r*-neighborhood of o, where *r* is a user-specified **distance threshold**
- An object o is an outlier if most (taking π as a **fraction threshold**) of the objects in D are far away from o, i.e., not in the r-neighborhood of o
- An object o is a DB(r, π) outlier if $\frac{\|\{o' | dist(o, o') \leq r\}\|}{\|D\|} \leq \pi$
- Equivalently, one can check the distance between *o* and its *k*-th nearest neighbor o_k, where $k = \lceil \pi \|D\| \rceil$. *o* is an outlier if dist(o, o_k) > r
- Efficient computation: Nested loop algorithm
  - For any object o_i, calculate its distance from other objects, and count the # of other objects in the r-neighborhood.
  - If π·n other objects are within r distance, terminate the inner loop
  - Otherwise, o_i is a DB(r, π) outlier
- Efficiency: Actually CPU time is not $O(n^2)$ but linear to the data set size since for most non-outlier objects, the inner loop terminates early

Distance-Based Outlier Detection: A Grid-Based Method

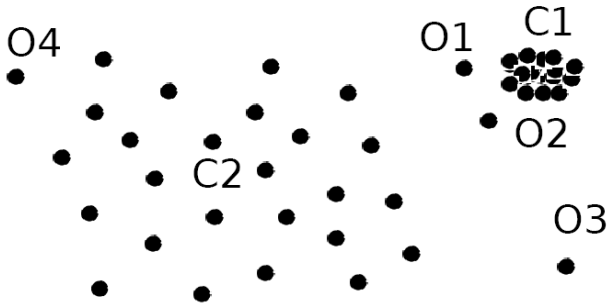
- Why efficiency is still a concern? When the complete set of objects cannot be held into main memory, cost I/O swapping
- The major cost: (1) each object tests against the whole data set, why not only its close neighbor? (2) check objects one by one, why not group by group?
- Grid-based method (CELL): Data space is partitioned into a multi-D grid. Each cell is a hyper cube with diagonal length r/2
- Pruning using the level-1 & level 2 cell properties:
  - For any possible point x in cell C and any possible point y in a level-1 cell, dist(x,y) ≤ r
  - For any possible point x in cell C and any point y such that dist(x,y) ≥ r, y is in a level-2 cell



- Thus we only need to check the objects that cannot be pruned, and even for such an object o, only need to compute the distance between o and the objects in the level-2 cells (since beyond level-2, the distance from o is more than r)

Density-Based Outlier Detection

- Local outliers: Outliers comparing to their local neighborhoods, instead of the global data distribution
- In Fig., o_1 and o2 are local outliers to C_1, o_3 is a global outlier, but o_4 is not an outlier. However, proximity-based clustering cannot find o_1 and o_2 are outlier (e.g., comparing with O_4).



- Intuition (density-based outlier detection): The density around an outlier object is significantly different from the density around its neighbors
- Method: Use the relative density of an object against its neighbors as the indicator of the degree of the object being outliers
- *k-distance* of an object o, $dist_k(o)$: distance between o and its k-th NN
- *k-distance neighborhood* of o, $N_k(o) = \{o' \mid o' \text{ in } D, dist(o, o') \leq dist_k(o)\}$
  - $N_k(o)$ could be bigger than k since multiple objects may have identical distance to o

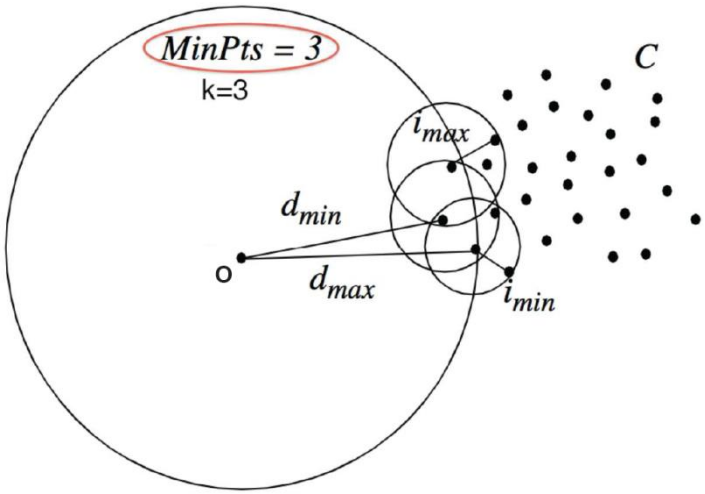
Local Outlier Factor: LOF

- Reachability distance from *o'* to *o*:

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

  - where k is a user-specified parameter
- Local reachability density of *o*:

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$



- LOF (Local outlier factor) of an object o is the average of the ratio of local reachability of *o* and those of *o*'s k-nearest neighbors

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)$$

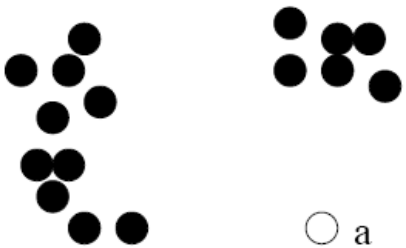
- The lower the local reachability density of o, and the higher the local reachability density of the kNN of o, the higher LOF
- This captures a local outlier whose local density is relatively low comparing to the local densities of its kNN

Chapter 12. Outlier Analysis

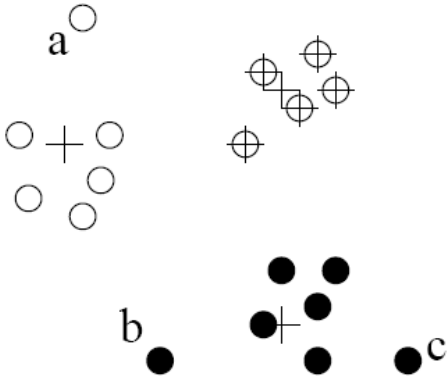
- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

Clustering-Based Outlier Detection (1 & 2):
Not belong to any cluster, or far from the closest one

- An object is an outlier if (1) it does not belong to any cluster, (2) there is a large distance between the object and its closest cluster , or (3) it belongs to a small or sparse cluster
- Case I: Not belong to any cluster
  - Identify animals not part of a flock: Using a density-based clustering method such as DBSCAN

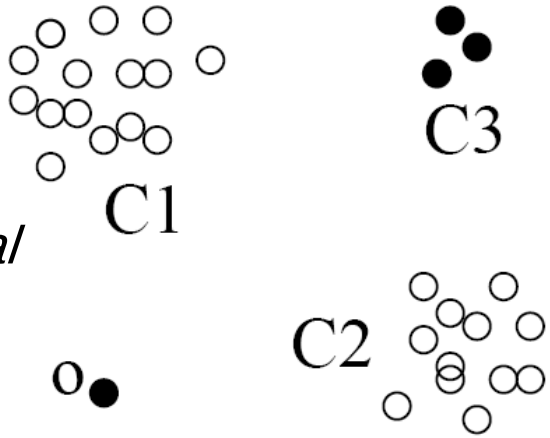


- Case 2: Far from its closest cluster
  - Using k-means, partition data points of into clusters
  - For each object o, assign an outlier score based on its distance from its closest center
    - If dist(o, c_o)/avg_dist(c_o) is large, likely an outlier



- Ex. Intrusion detection: Consider the similarity between data points and the clusters in a training data set
  - Use a training set to find patterns of “normal” data, e.g., frequent itemsets in each segment, and cluster similar connections into groups
  - Compare new data points with the clusters mined—Outliers are possible attacks

Clustering-Based Outlier Detection (3): Detecting Outliers in Small Clusters



- *FindCBLOF:* Detect outliers in small clusters
  - Find clusters, and sort them in decreasing size
  - To each data point, assign a *cluster-based local outlier factor* (CBLOF):
  - If obj p belongs to a large cluster, CBLOF = cluster_size X similarity between p and cluster
  - If p belongs to a small one, CBLOF = cluster size X similarity betw. p and the closest large cluster

- Ex. In the figure, o is outlier since its closest large cluster is C_1, but the similarity between o and C_1 is small. For any point in C_3, its closest large cluster is C_2 but its similarity from C_2 is low, plus $|C_3| = 3$ is small

Clustering-Based Method: Strength and Weakness

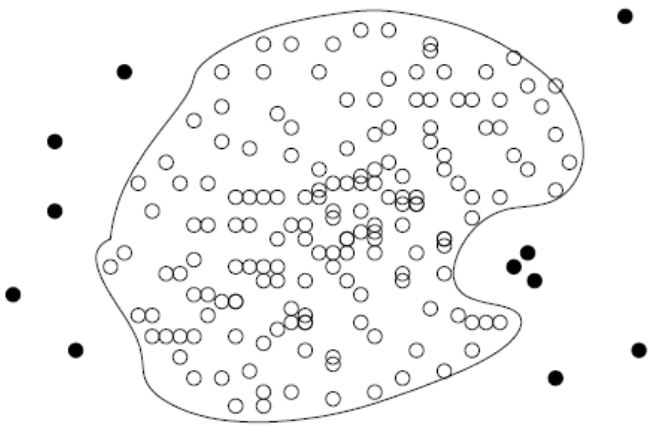
- Strength
  - Detect outliers without requiring any labeled data
  - Work for many types of data
  - Clusters can be regarded as summaries of the data
  - Once the cluster are obtained, need only compare any object against the clusters to determine whether it is an outlier (fast)
- Weakness
  - Effectiveness depends highly on the clustering method used—they may not be optimized for outlier detection
  - High computational cost: Need to first find clusters
  - A method to reduce the cost: Fixed-width clustering
    - A point is assigned to a cluster if the center of the cluster is within a pre-defined distance threshold from the point
    - If a point cannot be assigned to any existing cluster, a new cluster is created and the distance threshold may be learned from the training data under certain conditions

Chapter 12. Outlier Analysis

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

Classification-Based Method I: One-Class Model

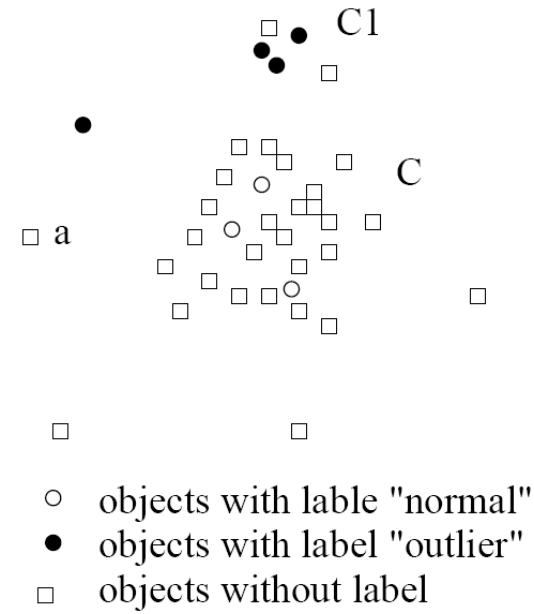
- Idea: Train a classification model that can distinguish "normal" data from outliers
- A brute-force approach: Consider a training set that contains samples labeled as "normal" and others labeled as "outlier"
  - But, the training set is typically heavily biased: # of "normal" samples likely far exceeds # of outlier samples
  - Cannot detect unseen anomaly



- One-class model: A classifier is built to describe only the normal class.
  - Learn the decision boundary of the normal class using classification methods such as SVM
  - Any samples that do not belong to the normal class (not within the decision boundary) are declared as outliers
  - Adv: can detect new outliers that may not appear close to any outlier objects in the training set
  - Extension: Normal objects may belong to multiple classes

Classification-Based Method II: Semi-Supervised Learning

- Semi-supervised learning: Combining classification-based and clustering-based methods
- Method
  - Using a clustering-based approach, find a large cluster, C, and a small cluster, C_1
  - Since some objects in C carry the label "normal", treat all objects in C as normal
  - Use the one-class model of this cluster to identify normal objects in outlier detection
  - Since some objects in cluster C_1 carry the label "outlier", declare all objects in C_1 as outliers
  - Any object that does not fall into the model for C (such as *a*) is considered an outlier as well



- Comments on classification-based outlier detection methods
  - Strength: Outlier detection is fast
  - Bottleneck: Quality heavily depends on the availability and quality of the training set, but often difficult to obtain representative and high-quality training data

Chapter 12. Outlier Analysis

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

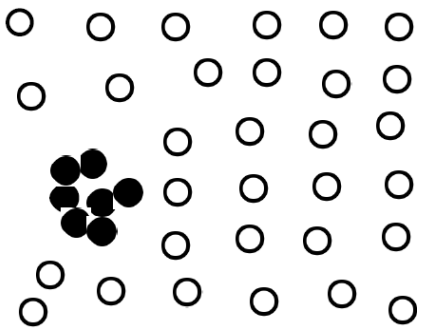
Mining Contextual Outliers I: Transform into Conventional Outlier Detection

- If the contexts can be clearly identified, transform it to conventional outlier detection
  1. Identify the context of the object using the contextual attributes
  2. Calculate the outlier score for the object in the context using a conventional outlier detection method
- Ex. Detect outlier customers in the context of customer groups
  - Contextual attributes: *age group, postal code*
  - Behavioral attributes: # *of trans/yr*, *annual total trans. amount*
- Steps: (1) locate c's context, (2) compare c with the other customers in the same group, and (3) use a conventional outlier detection method
- If the context contains very few customers, generalize contexts
  - Ex. Learn a mixture model U on the contextual attributes, and another mixture model V of the data on the behavior attributes
  - Learn a mapping $p(V_i|U_j)$: the probability that a data object o belonging to cluster U_j on the contextual attributes is generated by cluster Vi on the behavior attributes
  - Outlier score: $$S(\boldsymbol{o}) = \sum_{U_j} p(\boldsymbol{o} \in U_j) \sum_{V_i} p(\boldsymbol{o} \in V_i) p(V_i|U_j)$$

Mining Contextual Outliers II: Modeling Normal Behavior with Respect to Contexts

- In some applications, one cannot clearly partition the data into contexts
  - Ex. if a customer suddenly purchased a product that is unrelated to those she recently browsed, it is unclear how many products browsed earlier should be considered as the context
- Model the "normal" behavior with respect to contexts
  - Using a training data set, train a model that predicts the expected behavior attribute values with respect to the contextual attribute values
  - An object is a contextual outlier if its behavior attribute values significantly deviate from the values predicted by the model
- Using a prediction model that links the contexts and behavior, these methods avoid the explicit identification of specific contexts
- Methods: A number of classification and prediction techniques can be used to build such models, such as regression, Markov Models, and Finite State Automaton

Mining Collective Outliers I: On the Set of "Structured Objects"



- Collective outlier if objects as a group deviate significantly from the entire data
- Need to examine the *structure* of the data set, i.e, the relationships between multiple data objects
- Each of these structures is inherent to its respective type of data
  - For temporal data (such as time series and sequences), we explore the structures formed by time, which occur in segments of the time series or subsequences
  - For spatial data, explore local areas
  - For graph and network data, we explore subgraphs
- Difference from the contextual outlier detection: the structures are often not explicitly defined, and have to be discovered as part of the outlier detection process.
- Collective outlier detection methods: two categories
  - Reduce the problem to conventional outlier detection
    - Identify *structure units*, treat each structure unit (e.g., subsequence, time series segment, local area, or subgraph) as a data object, and extract features
    - Then outlier detection on the set of "structured objects" constructed as such using the extracted features

Mining Collective Outliers II: Direct Modeling of the Expected Behavior of Structure Units

- Models the expected behavior of structure units directly
- Ex. 1. Detect collective outliers in online social network of customers
  - Treat each possible subgraph of the network as a structure unit
  - Collective outlier: An *outlier subgraph* in the social network
    - Small subgraphs that are of very low frequency
    - Large subgraphs that are surprisingly frequent
- Ex. 2. Detect collective outliers in temporal sequences
  - Learn a Markov model from the sequences
  - A subsequence can then be declared as a collective outlier if it significantly deviates from the model
- Collective outlier detection is subtle due to the challenge of exploring the structures in data
  - The exploration typically uses heuristics, and thus may be application dependent
  - The computational cost is often high due to the sophisticated mining process

Chapter 12. Outlier Analysis

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

Challenges for Outlier Detection in High-Dimensional Data

- Interpretation of outliers
  - Detecting outliers without saying why they are outliers is not very useful in high-D due to many features (or dimensions) are involved in a high-dimensional data set
  - E.g., which subspaces that manifest the outliers or an assessment regarding the “outlier-ness” of the objects
- Data sparsity
  - Data in high-D spaces are often sparse
  - The distance between objects becomes heavily dominated by noise as the dimensionality increases
- Data subspaces
  - Adaptive to the subspaces signifying the outliers
  - Capturing the local behavior of data
- Scalable with respect to dimensionality
  - # of subspaces increases exponentially

Approach I: Extending Conventional Outlier Detection

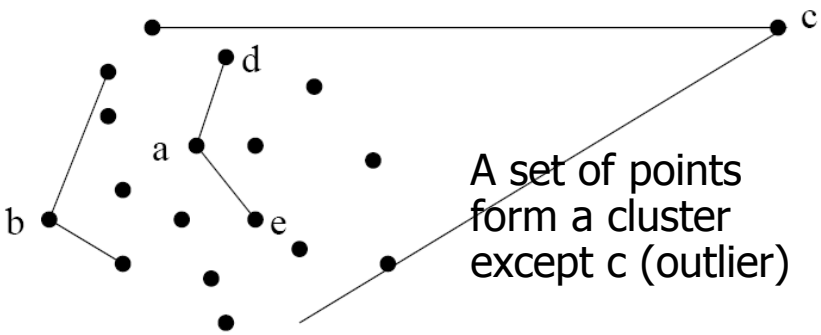
- Method 1: Detect outliers in the full space, e.g., HilOut Algorithm
  - Find distance-based outliers, but use the ranks of distance instead of the absolute distance in outlier detection
  - For each object *o*, find its k-nearest neighbors: $nn_1(o), \ldots, nn_k(o)$
  - The weight of object o: $$w(\boldsymbol{o}) = \sum_{i=1}^{k} dist(\boldsymbol{o}, nn_i(\boldsymbol{o}))$$
  - All objects are ranked in weight-descending order
  - Top-*l* objects in weight are output as outliers (*l*: user-specified parm)
  - Employ space-filling curves for approximation: scalable in both time and space w.r.t. data size and dimensionality
- Method 2: Dimensionality reduction
  - Works only when in lower-dimensionality, normal instances can still be distinguished from outliers
  - PCA: Heuristically, the principal components with low variance are preferred because, on such dimensions, normal objects are likely close to each other and outliers often deviate from the majority

Approach II: Finding Outliers in Subspaces

- Extending conventional outlier detection: Hard for outlier interpretation
- Find outliers in much lower dimensional subspaces: easy to interpret *why* and *to what extent* the object is an outlier
  - E.g., find outlier customers in certain subspace: *average transaction amount >> avg.* and *purchase frequency* << avg.
- Ex. A grid-based subspace outlier detection method
  - Project data onto various subspaces to find an area whose density is much lower than average
  - Discretize the data into a grid with φ equi-depth (why?) regions
  - Search for regions that are significantly sparse
    - Consider a k-d cube: k ranges on k dimensions, with n objects
    - If objects are independently distributed, the expected number of objects falling into a k-dimensional region is $(1/\varphi)^k n = f^k n$, the standard deviation is $\sqrt{f^k(1-f^k)n}$
    - The sparsity coefficient of cube C: $S(C) = \frac{n(C) - f^k n}{\sqrt{f^k(1-f^k)n}}$
    - If S(C) < 0, C contains less objects than expected
    - The more negative, the sparser C is and the more likely the objects in C are outliers in the subspace

Approach III: Modeling High-Dimensional Outliers

- Develop new models for high-dimensional outliers directly
- Avoid proximity measures and adopt new heuristics that do not deteriorate in high-dimensional data



- Ex. Angle-based outliers: Kriegel, Schubert, and Zimek [KSZ08]
- For each point o, examine the angle Δxoy for every pair of points x, y.
  - Point in the center (e.g., a), the angles formed differ widely
  - An outlier (e.g., c), angle variable is substantially smaller
- Use the variance of angles for a point to determine outlier
- Combine angles and distance to model outliers
  - Use the distance-weighted angle variance as the outlier score
  - Angle-based outlier factor (ABOF):

$$ABOF(\boldsymbol{o}) = VAR_{\boldsymbol{x},\boldsymbol{y}\in D,\boldsymbol{x}\neq\boldsymbol{o},\boldsymbol{y}\neq\boldsymbol{o}} \frac{\langle \overrightarrow{\boldsymbol{ox}}, \overrightarrow{\boldsymbol{oy}} \rangle}{dist(\boldsymbol{o},\boldsymbol{x})^2 dist(\boldsymbol{o},\boldsymbol{y})^2}$$

  - Efficient approximation computation method is developed
  - It can be generalized to handle arbitrary types of data

Chapter 12. Outlier Analysis

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

Summary

- Types of outliers
  - global, contextual & collective outliers
- Outlier detection
  - supervised, semi-supervised, or unsupervised
- Statistical (or model-based) approaches
- Proximity-base approaches
- Clustering-base approaches
- Classification approaches
- Mining contextual and collective outliers
- Outlier detection in high dimensional data

References (I)

- B. Abraham and G.E.P. Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 66:229–248, 1979.
- M. Agyemang, K. Barker, and R. Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intell. Data Anal.*, 10:521–538, 2006.
- F. J. Anscombe and I. Guttman. Rejection of outliers. *Technometrics*, 2:123–147, 1960.
- D. Agarwal. Detecting anomalies in cross-classified streams: a bayesian approach. *Knowl. Inf. Syst.*, 11:29–44, 2006.
- F. Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. *TKDE,* 2005.
- C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. *SIGMOD'01*
- R.J. Beckman and R.D. Cook. Outlier...s. *Technometrics*, 25:119–149, 1983.
- I. Ben-Gal. Outlier detection. In *Maimon O. and Rockach L. (eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic, 2005.
- M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. *SIGMOD'00*
- D. Barbar´a, Y. Li, J. Couto, J.-L. Lin, and S. Jajodia. Bootstrapping a data mining intrusion detection system. *SAC'03*
- Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris. A comparative study for outlier detection techniques in data mining. *IEEE Conf. on Cybernetics and Intelligent Systems*, 2006.
- S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. *KDD'03*
- D. Barbara, N. Wu, and S. Jajodia. Detecting novel network intrusion using bayesian estimators. *SDM'01*
- V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58, 2009.
- D. Dasgupta and N.S. Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. In *CEC'02*

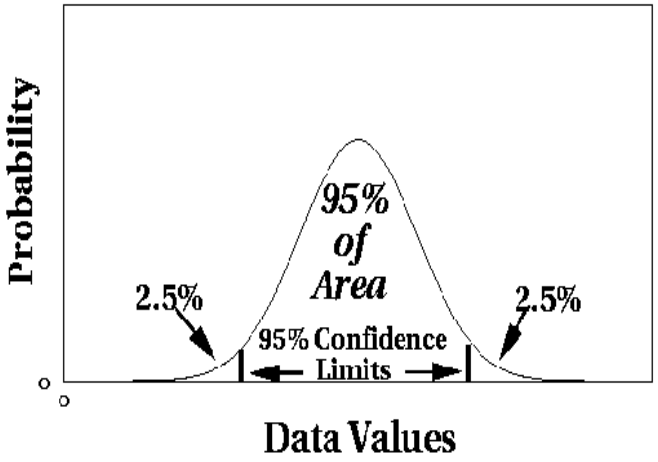
References (2)

- E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Proc. 2002 Int. Conf. of Data Mining for Security Applications*, 2002.
- E. Eskin. Anomaly detection over noisy data using learned probability distributions. *ICML'00*
- T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291–316, 1997.
- V. J. Hodge and J. Austin. A survey of outlier detection methdologies. *Artif. Intell. Rev.*, 22:85–126, 2004.
- D. M. Hawkins. *Identification of Outliers*. Chapman and Hall, London, 1980.
- Z. He, X. Xu, and S. Deng. Discovering cluster-based local outliers. *Pattern Recogn. Lett.*, 24, June, 2003.
- W. Jin, K. H. Tung, and J. Han. Mining top-n local outliers in large databases. *KDD'01*
- W. Jin, A. K. H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. *PAKDD'06*
- E. Knorr and R. Ng. A unified notion of outliers: Properties and computation. *KDD'97*
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. *VLDB'98*
- E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *VLDB J.*, 8:237–253, 2000.
- H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. *KDD'08*
- M. Markou and S. Singh. Novelty detection: A review—part 1: Statistical approaches. *Signal Process.*, 83:2481–2497, 2003.
- M. Markou and S. Singh. Novelty detection: A review—part 2: Neural network based approaches. *Signal Process.*, 83:2499–2521, 2003.
- C. C. Noble and D. J. Cook. Graph-based anomaly detection. *KDD'03*

References (3)

- S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. *ICDE'03*
- A. Patcha and J.-M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.*, 51, 2007.
- X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Trans. on Knowl. and Data Eng.*, 19, 2007.
- Y. Tao, X. Xiao, and S. Zhou. Mining distance-based outliers from large databases in any metric space. *KDD'06*
- N. Ye and Q. Chen. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International*, 17:105–112, 2001.
- B.-K. Yi, N. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. *ICDE'00*

Outlier Discovery: Statistical Approaches



Assume a model underlying distribution that generates data set (e.g. normal distribution)

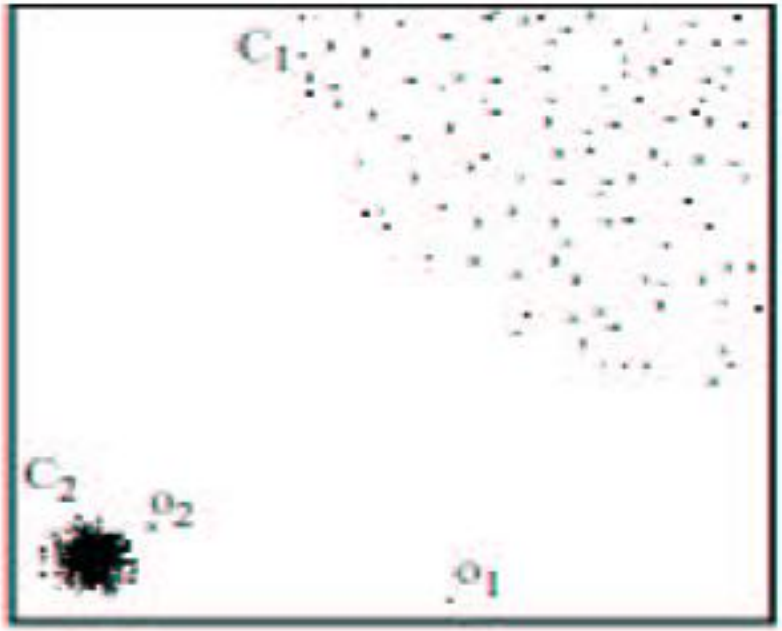
- Use discordancy tests depending on
  - data distribution
  - distribution parameter (e.g., mean, variance)
  - number of expected outliers
- Drawbacks
  - most tests are for single attribute
  - In many cases, data distribution may not be known

Outlier Discovery: Distance-Based Approach

- Introduced to counter the main limitations imposed by statistical methods
  - We need multi-dimensional analysis without knowing data distribution
- Distance-based outlier: A DB(p, D)-outlier is an object O in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than D from O
- Algorithms for mining distance-based outliers [Knorr & Ng, VLDB'98]
  - Index-based algorithm
  - Nested-loop algorithm
  - Cell-based algorithm

Density-Based Local Outlier Detection

- M. M. Breunig, H.-P. Kriegel, R. Ng, J. Sander. LOF: Identifying Density-Based Local Outliers. SIGMOD 2000.
- Distance-based outlier detection is based on global distance distribution
- It encounters difficulties to identify outliers if data is not uniformly distributed
- Ex. C_1 contains 400 loosely distributed points, C_2 has 100 tightly condensed points, 2 outlier points o_1, o_2
- Distance-based method cannot identify o_2 as an outlier



- Need the concept of local outlier
- Local outlier factor (LOF)
  - Assume outlier is not crisp
  - Each point has a LOF

Outlier Discovery: Deviation-Based Approach

- Identifies outliers by examining the main characteristics of objects in a group
- Objects that “deviate” from this description are considered outliers
- Sequential exception technique
  - simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects
- OLAP data cube technique
  - uses data cubes to identify regions of anomalies in large multidimensional data

References (1)

- B. Abraham and G.E.P. Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 1979.
- Malik Agyemang, Ken Barker, and Rada Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intell. Data Anal.*, 2006.
- Deepak Agarwal. Detecting anomalies in cross-classied streams: a bayesian approach. *Knowl. Inf. Syst.*, 2006.
- C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. *SIGMOD'01*.
- M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Optics-of: Identifying local outliers. *PKDD '99*
- M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. *SIGMOD'00*.
- V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 2009.
- D. Dasgupta and N.S. Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. *Computational Intelligence,* 2002.
- E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Proc. 2002 Int. Conf. of Data Mining for Security Applications*, 2002.
- E. Eskin. Anomaly detection over noisy data using learned probability distributions. *ICML'00*.
- T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1997.
- R. Fujimaki, T. Yairi, and K. Machida. An approach to spacecraft anomaly detection problem using kernel feature space. *KDD '05*
- F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 1969.

References (2)

- V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 2004.
- Douglas M Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.
- P. S. Horn, L. Feng, Y. Li, and A. J. Pesce. Effect of Outliers and Nonhealthy Individuals on Reference Interval Estimation. *Clin Chem*, 2001.
- W. Jin, A. K. H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. *PAKDD'06*
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98
- M. Markou and S. Singh.. Novelty detection: a review| part 1: statistical approaches. *Signal Process.*, 83(12), 2003.
- M. Markou and S. Singh. Novelty detection: a review| part 2: neural network based approaches. *Signal Process.*, 83(12), 2003.
- S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. *ICDE'03*.
- A. Patcha and J.-M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.*, 51(12):3448{3470, 2007.
- W. Stefansky. Rejecting outliers in factorial designs. *Technometrics*, 14(2):469{479, 1972.
- X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Trans. on Knowl. and Data Eng.*, 19(5):631{645, 2007.
- Y. Tao, X. Xiao, and S. Zhou. Mining distance-based outliers from large databases in any metric space. *KDD '06:*
- N. Ye and Q. Chen. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International*, 2001.