
Data Mining:

Concepts and Techniques


(3rd ed.)

unit -5

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

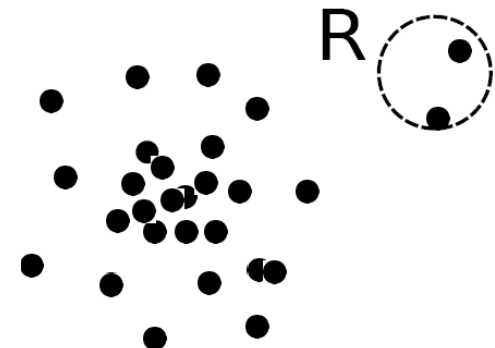
©2011 Han, Kamber & Pei. All rights reserved.

Outlier Analysis

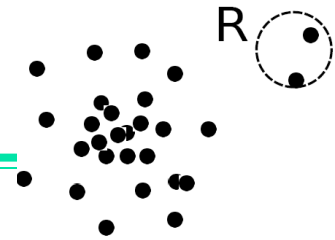
- Outlier and Outlier Analysis 
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary

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)



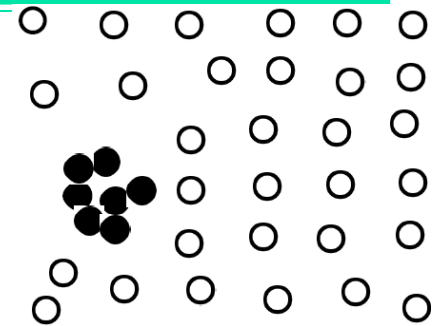
Global Outlier

- 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?

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




Collective Outlier

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 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 whether user-labeled examples of outliers can be obtained:
 - Supervised, semi-supervised vs. unsupervised methods
 - Based on assumptions about normal data and outliers:
 - 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

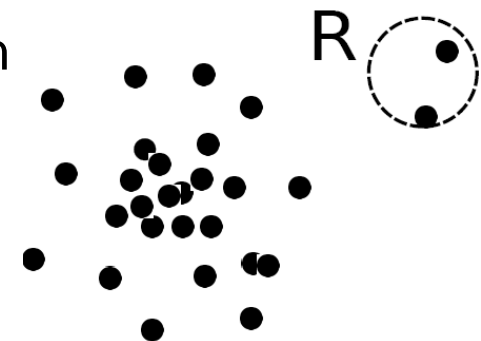
- Assume the normal objects are somewhat “clustered” into multiple groups, each having some distinct features
- An outlier is expected to be far away from any groups of normal objects
- Weakness: Cannot detect collective outlier effectively
 - Normal objects may not share any strong patterns, but the 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

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

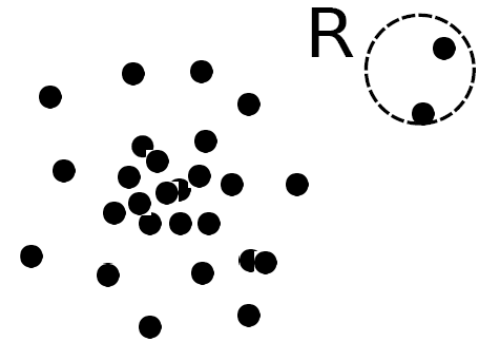
Outlier Detection (1): Statistical Methods

- 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



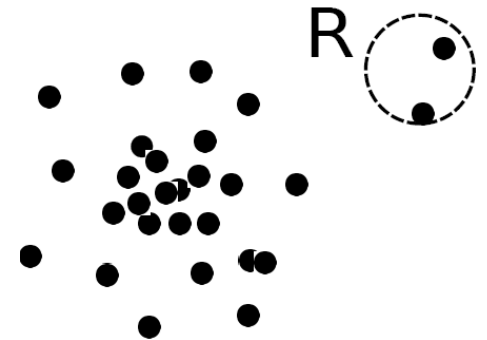
Outlier Detection (2): Proximity-Based Methods

- 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 (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



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

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

- For the above data with $n = 10$, we have $\hat{\mu} = 28.61$ $\hat{\sigma} = \sqrt{2.29} = 1.51$
- Then $(24 - 28.61) / 1.51 = -3.04 < -3$, 24 is an outlier since $\mu \pm 3\sigma$ region contains 99.7% 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

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

where $t_{\alpha/(2N), N-2}^2$ is the value taken by a t-distribution at a significance level of $\alpha/(2N)$, 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 $\text{MDist}(o, \bar{o}) = (o - \bar{o})^T S^{-1}(o - \bar{o})$ 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

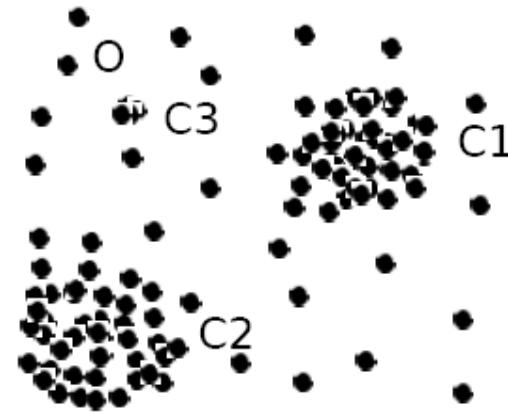
Parametric Methods III: Using Mixture of Parametric Distributions

- 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

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

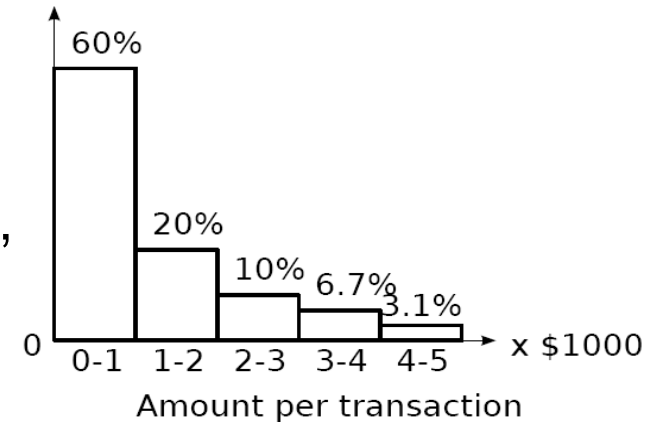
where f_{θ_1} and f_{θ_2} are the probability density functions of θ_1 and θ_2

- Then use EM algorithm to learn the parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$ 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.



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 $DB(r, \pi)$ 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 $\pi \cdot n$ other objects are within r distance, terminate the inner loop
 - Otherwise, o_i is a $DB(r, \pi)$ 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

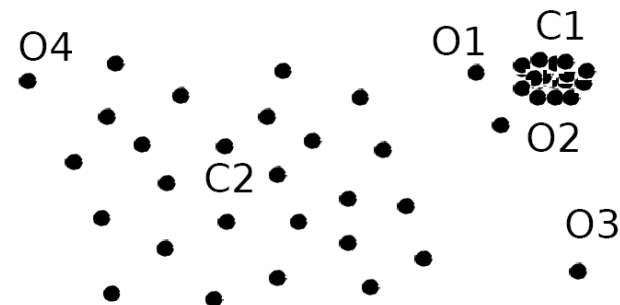
Distance-Based Outlier Detection: A Grid-Based Method

- 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, $\text{dist}(x,y) \leq r$
 - For any possible point x in cell C and any point y such that $\text{dist}(x,y) \geq r$, 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)

2	2	2	2	2	2	2
2	2	2	2	2	2	2
2	2	1	1	1	2	2
2	2	1	C	1	2	2
2	2	1	1	1	2	2
2	2	2	2	2	2	2
2	2	2	2	2	2	2

Density-Based Outlier Detection

- Local outliers: Outliers comparing to their local neighborhoods, instead of the global data distribution
- In Fig., o_1 and o_2 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 , $\text{dist}_k(o)$: distance between o and its k -th NN
- k -distance neighborhood of o , $N_k(o) = \{o' \mid o' \text{ in } D, \text{dist}(o, o') \leq \text{dist}_k(o)\}$
 - $N_k(o)$ 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

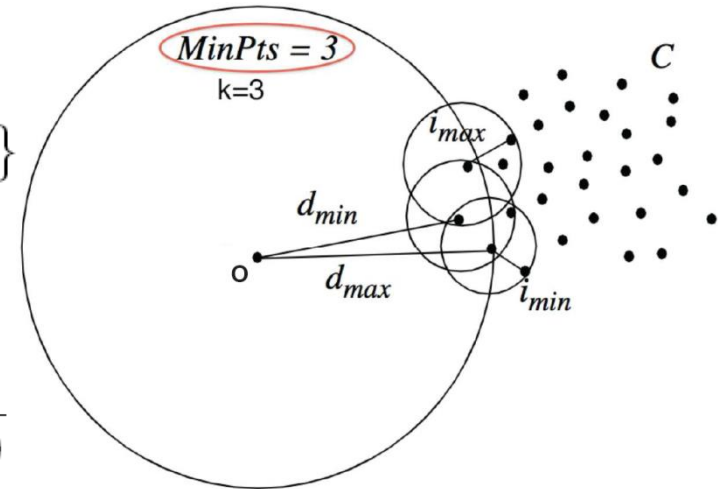
- Local reachability density of o :

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


- 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 k NN of o , the higher LOF
- This captures a local outlier whose local density is relatively low comparing to the local densities of its k NN



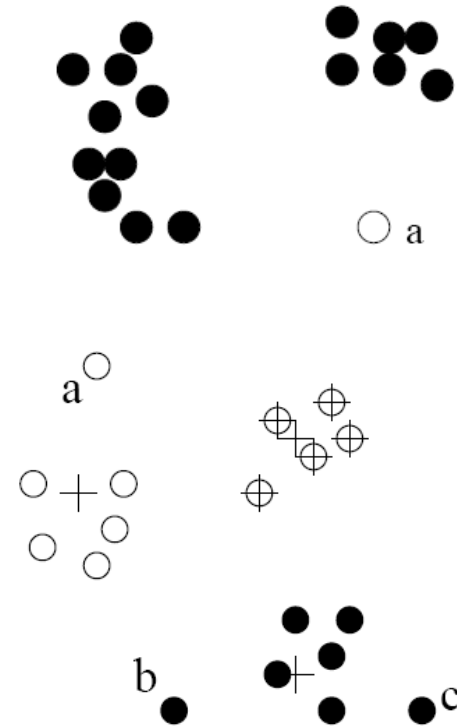
Outlier Analysis

- 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):

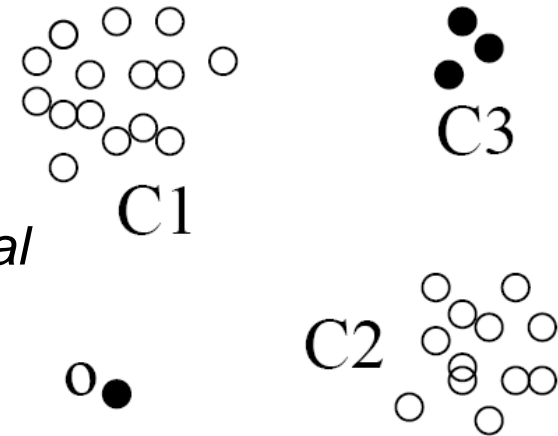
Not belong to any cluster, or far from the closest one

- 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 1: 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 $\text{dist}(o, c_o) / \text{avg_dist}(c_o)$ 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, $CBLOF = \text{cluster_size} \times \text{similarity between } p \text{ and cluster}$
 - If p belongs to a small one, $CBLOF = \text{cluster size} \times \text{similarity betw. } p \text{ and the closest large cluster}$
- 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

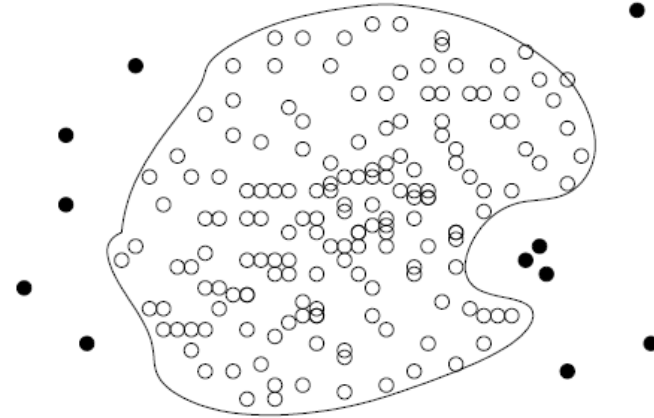
- Strength
 - Detect outliers without requiring any labeled data
 - Work for many types of data
 - Clusters can be regarded as summaries of the data
 - Once the cluster are obtained, need only compare any object against the clusters to determine whether it is an outlier (fast)
- Weakness
 - Effectiveness 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

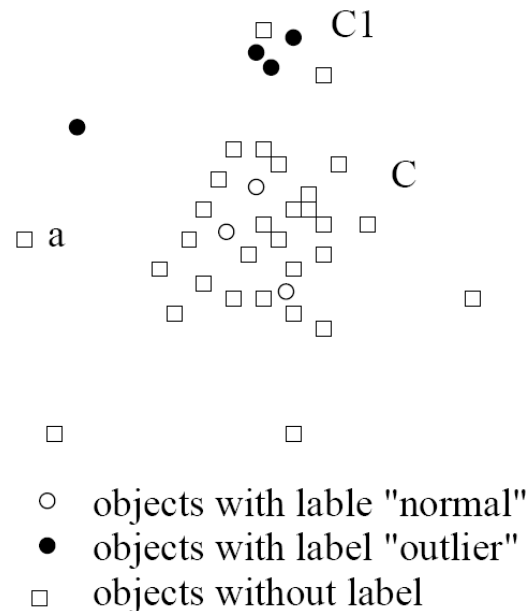
Classification-Based Method I: One-Class Model

- 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

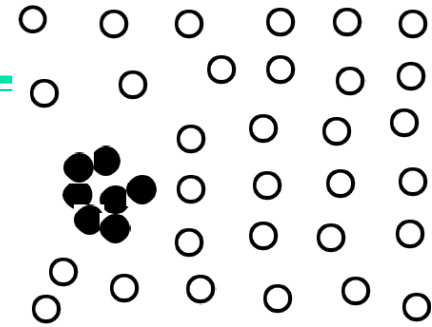
- If the contexts can be clearly identified, transform it to conventional outlier detection
 1. Identify the context of the object using the contextual attributes
 2. Calculate the outlier score for the object in the context using a conventional outlier detection method
- Ex. Detect outlier customers 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 V_i on the behavior attributes
 - Outlier score:
$$S(o) = \sum_{U_j} p(o \in U_j) \sum_{V_i} p(o \in V_i) p(V_i|U_j)$$

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

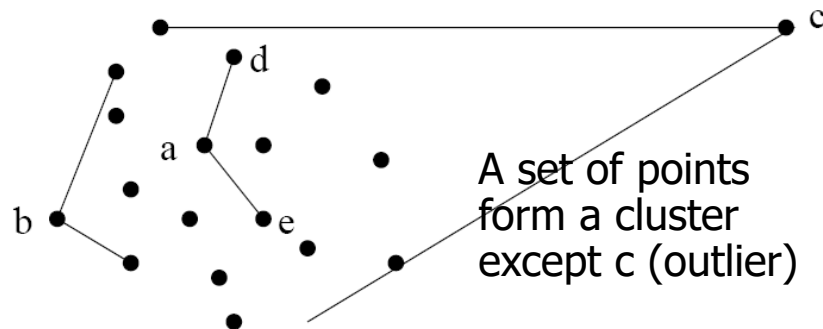
- Method 1: Detect outliers in the full space, e.g., HilOut Algorithm
 - Find distance-based outliers, but use the ranks of distance instead of the absolute distance in outlier detection
 - For each object o , find its k -nearest neighbors: $nn_1(o), \dots, nn_k(o)$
 - The weight of object o :
$$w(o) = \sum_{i=1}^k dist(o, nn_i(o))$$
 - 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 $(1/\phi)^k n = f^k n$, 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 $\angle xoy$ 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):



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

- 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)

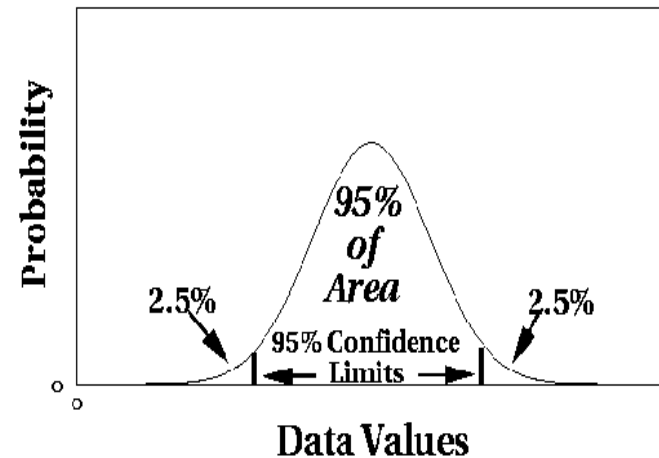
- E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Proc. 2002 Int. Conf. of Data Mining for Security Applications*, 2002.
- E. Eskin. Anomaly detection over noisy data using learned probability 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 methodologies. *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*

Un-Used Slides

Outlier Discovery: Statistical Approaches



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

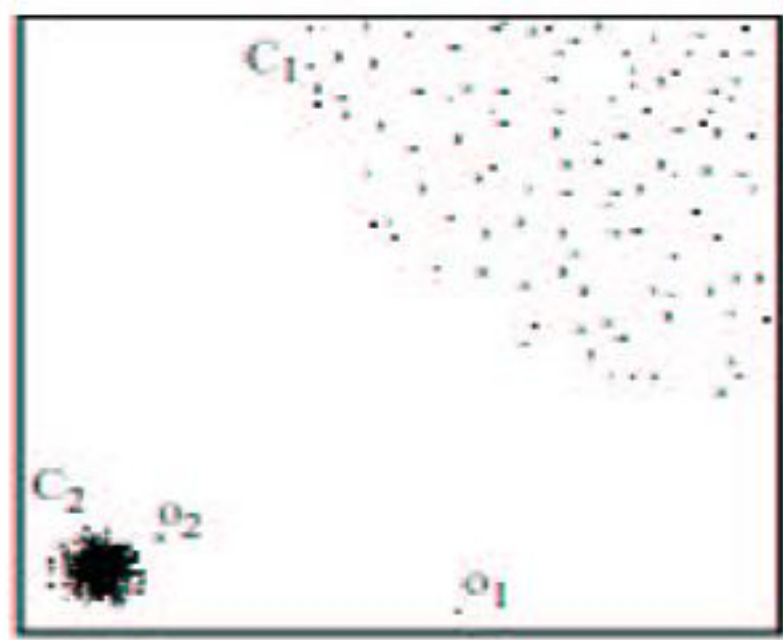
- Use discordancy tests depending on
 - data distribution
 - distribution parameter (e.g., mean, variance)
 - number of expected outliers
- Drawbacks
 - most tests are for single attribute
 - In many cases, data distribution may not be known

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-classified 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.

Data Mining:

Concepts and Techniques

(3rd ed.)

unit 5

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

©2011 Han, Kamber & Pei. All rights reserved.

Other Methodologies of Data Mining

- Statistical Data Mining 
- Views on Data Mining Foundations
- Visual and Audio Data Mining

Major Statistical Data Mining Methods

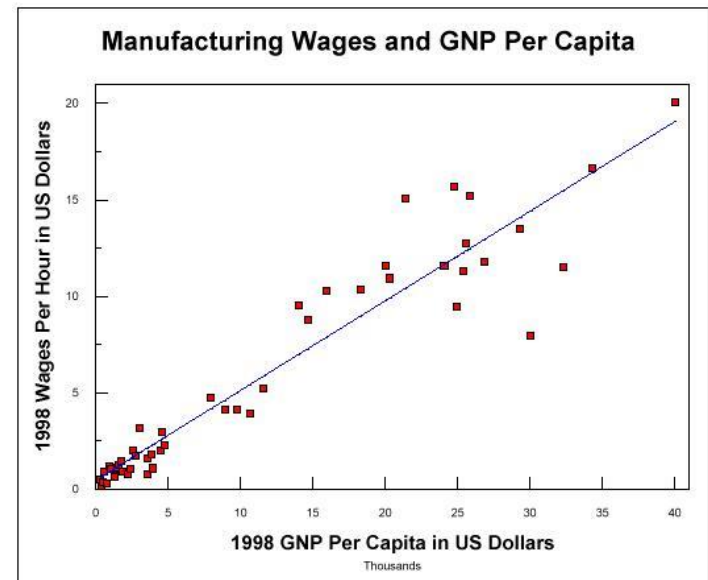
- Regression
- Generalized Linear Model
- Analysis of Variance
- Mixed-Effect Models
- Factor Analysis
- Discriminant Analysis
- Survival Analysis

Statistical Data Mining (1)

- There are many well-established statistical techniques for data analysis, particularly for numeric data
 - applied extensively to data from scientific experiments and data from economics and the social sciences

- **Regression**

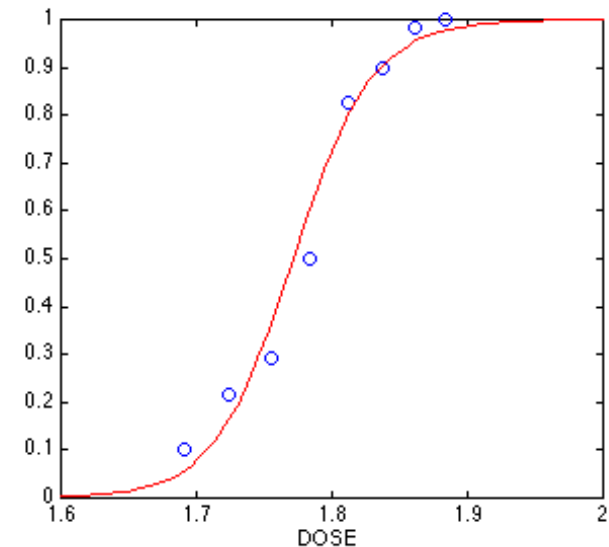
- predict the value of a **response** (dependent) variable from one or more **predictor** (independent) variables where the variables are numeric
 - forms of regression: linear, multiple, weighted, polynomial, nonparametric, and robust



Scientific and Statistical Data Mining (2)

■ Generalized linear models

- allow a categorical response variable (or some transformation of it) to be related to a set of predictor variables
- similar to the modeling of a numeric response variable using linear regression
- include logistic regression and Poisson regression



■ Mixed-effect models

- For analyzing **grouped data**, i.e. data that can be classified according to one or more grouping variables
- Typically describe relationships between a response variable and some covariates in data grouped according to one or more factors

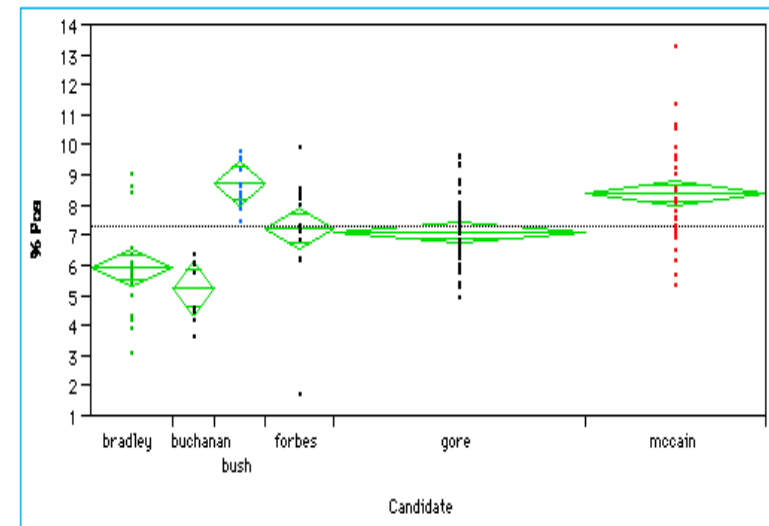
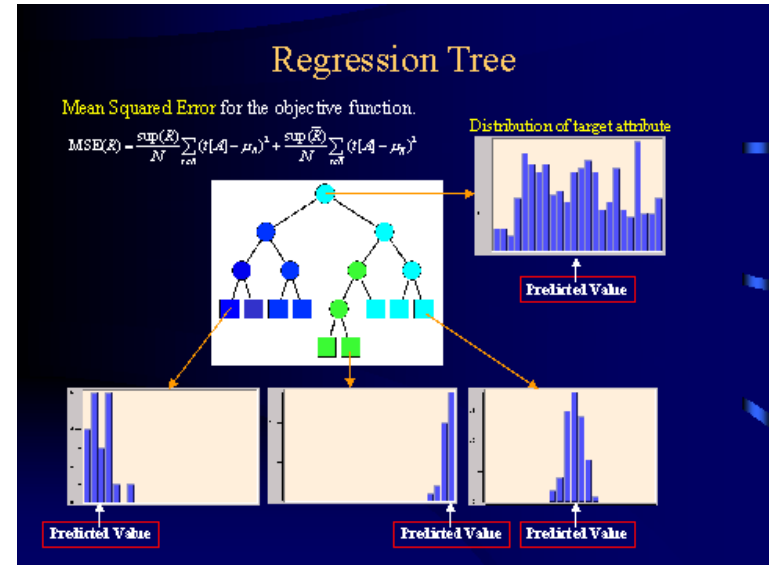
Scientific and Statistical Data Mining (3)

■ Regression trees

- Binary trees used for classification and prediction
- Similar to decision trees: Tests are performed at the internal nodes
- In a regression tree the mean of the objective attribute is computed and used as the predicted value

■ Analysis of variance

- Analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors)



Statistical Data Mining (4)

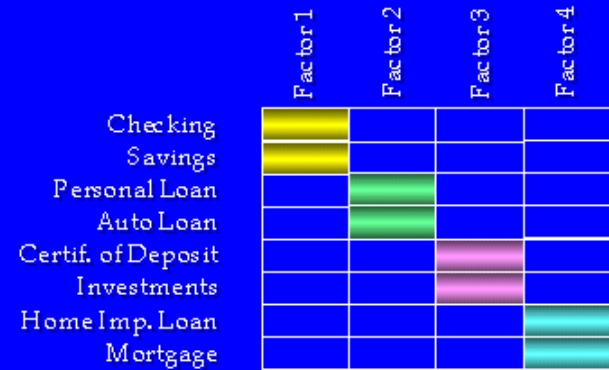
■ Factor analysis

- determine which variables are combined to generate a given factor
- e.g., for many psychiatric data, one can indirectly measure other quantities (such as test scores) that reflect the factor of interest

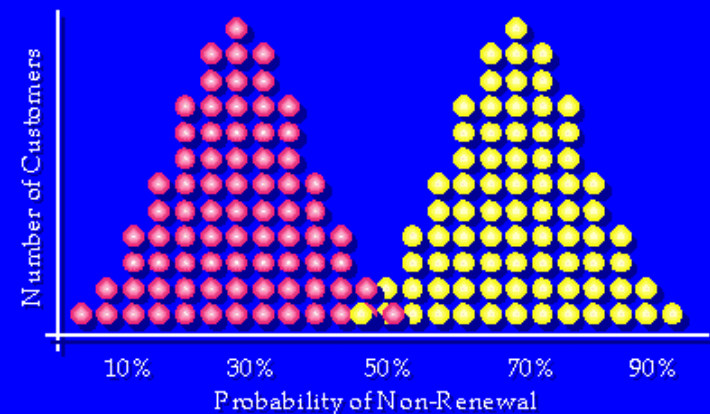
■ Discriminant analysis

- predict a categorical response variable, commonly used in social science
- Attempts to determine several discriminant functions (linear combinations of the independent variables) that discriminate among the groups defined by the response variable

Data Mining - Factor Analysis



Data Mining - Discriminant

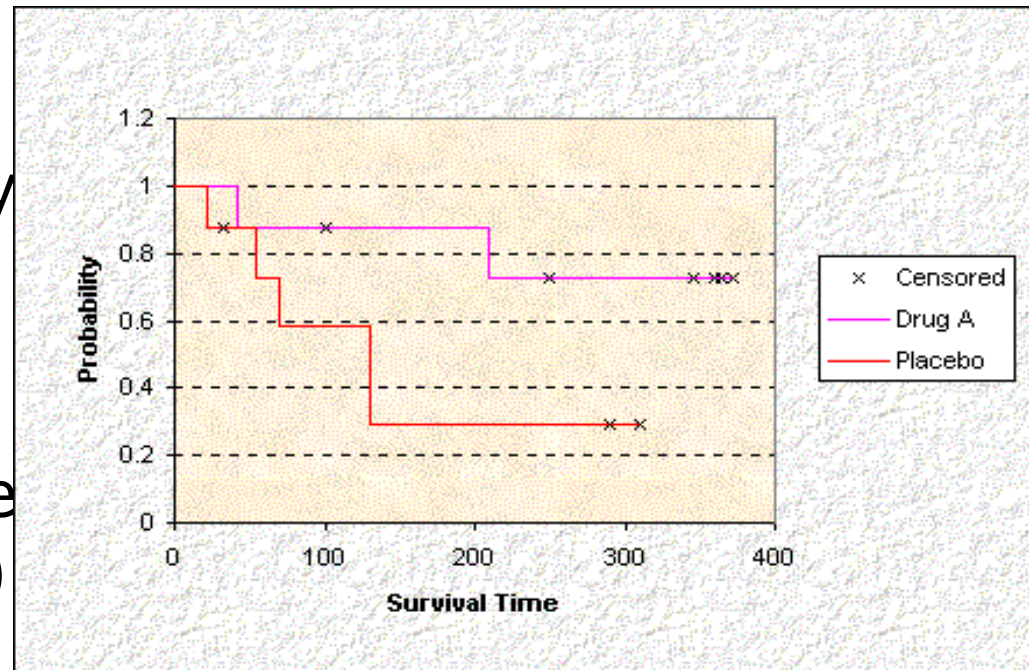


Statistical Data Mining (5)

- **Time series:** many methods such as autoregression, ARIMA (Autoregressive integrated moving-average modeling), long memory time-series modeling
- **Quality control:** displays group summary charts

- **Survival analysis**

- Predicts the probability that a patient undergoing a medical treatment would survive at least to time t (life span prediction)



Chapter 13: Data Mining Trends and Research Frontiers

- Mining Complex Types of Data
- Other Methodologies of Data Mining
- Data Mining Applications 
- Data Mining and Society
- Data Mining Trends
- Summary

Data Mining Applications

- Data mining: A young discipline with broad and diverse applications
 - There still exists a nontrivial gap between generic data mining methods and effective and scalable data mining tools for domain-specific applications
- Some application domains (briefly discussed here)
 - Data Mining for Financial data analysis
 - Data Mining for Retail and Telecommunication Industries
 - Data Mining in Science and Engineering
 - Data Mining for Intrusion Detection and Prevention
 - Data Mining and Recommender Systems

Data Mining for Financial Data Analysis (I)

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
- Design and construction of data warehouses for multidimensional data analysis and data mining
 - View the debt and revenue changes by month, by region, by sector, and by other factors
 - Access statistical information such as max, min, total, average, trend, etc.
- Loan payment prediction/consumer credit policy analysis
 - feature selection and attribute relevance ranking
 - Loan payment performance
 - Consumer credit rating

Data Mining for Financial Data Analysis (II)

- Classification and clustering of customers for targeted marketing
 - multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- Detection of money laundering and other financial crimes
 - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
 - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Data Mining for Retail & Telcomm. Industries (I)

- Retail industry: huge amounts of data on sales, customer shopping history, e-commerce, etc.
- Applications of retail data mining
 - Identify customer buying behaviors
 - Discover customer shopping patterns and trends
 - Improve the quality of customer service
 - Achieve better customer retention and satisfaction
 - Enhance goods consumption ratios
 - Design more effective goods transportation and distribution policies
- Telcomm. and many other industries: Share many similar goals and expectations of retail data mining

Data Mining Practice for Retail Industry

- Design and construction of data warehouses
- Multidimensional analysis of sales, customers, products, time, and region
- Analysis of the effectiveness of sales campaigns
- Customer retention: Analysis of customer loyalty
 - Use customer loyalty card information to register sequences of purchases of particular customers
 - Use sequential pattern mining to investigate changes in customer consumption or loyalty
 - Suggest adjustments on the pricing and variety of goods
- Product recommendation and cross-reference of items
- Fraudulent analysis and the identification of usual patterns
- Use of visualization tools in data analysis

Data Mining in Science and Engineering

- Data warehouses and data preprocessing
 - Resolving inconsistencies or incompatible data collected in diverse environments and different periods (e.g. eco-system studies)
- Mining complex data types
 - Spatiotemporal, biological, diverse semantics and relationships
- Graph-based and network-based mining
 - Links, relationships, data flow, etc.
- Visualization tools and domain-specific knowledge
- Other issues
 - Data mining in social sciences and social studies: text and social media
 - Data mining in computer science: monitoring systems, software bugs, network intrusion

Data Mining for Intrusion Detection and Prevention

- Majority of intrusion detection and prevention systems use
 - Signature-based detection: use signatures, attack patterns that are preconfigured and predetermined by domain experts
 - Anomaly-based detection: build profiles (models of normal behavior) and detect those that are substantially deviate from the profiles
- What data mining can help
 - New data mining algorithms for intrusion detection
 - Association, correlation, and discriminative pattern analysis help select and build discriminative classifiers
 - Analysis of stream data: outlier detection, clustering, model shifting
 - Distributed data mining
 - Visualization and querying tools

Data Mining and Recommender Systems

- Recommender systems: Personalization, making product recommendations that are likely to be of interest to a user
- Approaches: Content-based, collaborative, or their hybrid
 - Content-based: Recommends items that are similar to items the user preferred or queried in the past
 - Collaborative filtering: Consider a user's social environment, opinions of other customers who have similar tastes or preferences
- Data mining and recommender systems
 - Users $C \times$ items S : extract from known to unknown ratings to predict user-item combinations
 - Memory-based method often uses k-nearest neighbor approach
 - Model-based method uses a collection of ratings to learn a model (e.g., probabilistic models, clustering, Bayesian networks, etc.)
 - Hybrid approaches integrate both to improve performance (e.g., using ensemble)

Chapter 13: Data Mining Trends and Research Frontiers

- Mining Complex Types of Data
- Other Methodologies of Data Mining
- Data Mining Applications
- Data Mining and Society 
- Data Mining Trends
- Summary

Ubiquitous and Invisible Data Mining

- Ubiquitous Data Mining
 - Data mining is used everywhere, e.g., online shopping
 - Ex. Customer relationship management (CRM)
- Invisible Data Mining
 - Invisible: Data mining functions are built in daily life operations
 - Ex. Google search: Users may be unaware that they are examining results returned by data
 - Invisible data mining is highly desirable
 - Invisible mining needs to consider efficiency and scalability, user interaction, incorporation of background knowledge and visualization techniques, finding interesting patterns, real-time, ...
 - Further work: Integration of data mining into existing business and scientific technologies to provide domain-specific data mining tools

Privacy, Security and Social Impacts of Data Mining

- Many data mining applications do not touch personal data
 - E.g., meteorology, astronomy, geography, geology, biology, and other scientific and engineering data
- Many DM studies are on developing scalable algorithms to find general or statistically significant patterns, not touching individuals
- The real privacy concern: unconstrained access of individual records, especially privacy-sensitive information
- Method 1: Removing sensitive IDs associated with the data
- Method 2: Data security-enhancing methods
 - Multi-level security model: permit to access to only authorized level
 - Encryption: e.g., *blind signatures*, *biometric encryption*, and *anonymous databases* (personal information is encrypted and stored at different locations)
- Method 3: Privacy-preserving data mining methods

Privacy-Preserving Data Mining

- Privacy-preserving (privacy-enhanced or privacy-sensitive) mining:
 - Obtaining valid mining results without disclosing the underlying sensitive data values
 - Often needs trade-off between information loss and privacy
- Privacy-preserving data mining methods:
 - Randomization (e.g., perturbation): Add noise to the data in order to mask some attribute values of records
 - K-anonymity and l-diversity: Alter individual records so that they cannot be uniquely identified
 - k-anonymity: Any given record maps onto at least k other records
 - l-diversity: enforcing intra-group diversity of sensitive values
 - Distributed privacy preservation: Data partitioned and distributed either horizontally, vertically, or a combination of both
 - Downgrading the effectiveness of data mining: The output of data mining may violate privacy
 - Modify data or mining results, e.g., hiding some association rules or slightly distorting some classification models

Chapter 13: Data Mining Trends and Research Frontiers

- Mining Complex Types of Data
- Other Methodologies of Data Mining
- Data Mining Applications
- Data Mining and Society
- Data Mining Trends 
- Summary

Trends of Data Mining

- Application exploration: Dealing with application-specific problems
- Scalable and interactive data mining methods
- Integration of data mining with Web search engines, database systems, data warehouse systems and cloud computing systems
- Mining social and information networks
- Mining spatiotemporal, moving objects and cyber-physical systems
- Mining multimedia, text and web data
- Mining biological and biomedical data
- Data mining with software engineering and system engineering
- Visual and audio data mining
- Distributed data mining and real-time data stream mining
- Privacy protection and information security in data mining

Chapter 13: Data Mining Trends and Research Frontiers

- Mining Complex Types of Data
- Other Methodologies of Data Mining
- Data Mining Applications
- Data Mining and Society
- Data Mining Trends
- Summary 

Summary

- We present a high-level overview of **mining complex data types**
- **Statistical data mining** methods, such as regression, generalized linear models, analysis of variance, etc., are popularly adopted
- Researchers also try to build **theoretical foundations** for data mining
- **Visual/audio data mining** has been popular and effective
- **Application-based mining** integrates domain-specific knowledge with data analysis techniques and provide mission-specific solutions
- **Ubiquitous data mining** and **invisible data mining** are penetrating our data lives
- **Privacy and data security** are importance issues in data mining, and **privacy-preserving data mining** has been developed recently
- Our discussion on **trends in data mining** shows that data mining is a promising, young field, with great, strategic importance

References and Further Reading

- ❖ The books lists a lot of references for further reading. Here we only list a few books
 - E. Alpaydin. *Introduction to Machine Learning*, 2nd ed., MIT Press, 2011
 - S. Chakrabarti. *Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data*. Morgan Kaufmann, 2002
 - R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*, 2ed., Wiley-Interscience, 2000
 - D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.
 - U. Fayyad, G. Grinstein, and A. Wierse (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, 2001
 - J. Han, M. Kamber, J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd ed. 2011
 - T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer-Verlag, 2009
 - D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
 - B. Liu. *Web Data Mining*, Springer 2006.
 - T. M. Mitchell. *Machine Learning*, McGraw Hill, 1997
 - M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.
 - P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Wiley, 2005
 - I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 2nd ed. 2005