Chapter 6: Anomaly/Fraud Detection

Anomaly Detection

- Anomaly detection is a form of classification.
- Is the process to localize objects that are different from other objects (anomalies).
- The set of data points that are considerably different than the remainder of the data are anomalies/outliers.
- Anomaly detection is the process of detecting something unusual relative to something expected.
- The goal of anomaly detection is to identify cases that are unusual within data that is seemingly homogeneous.

Why is Anomaly Detection important?

- to detect problems
- to detect new phenomenon
- to discover unusual behavior in data
-

Examples of interesting application for Anomaly Detection

- Fraud Detection - looking for buying patterns different from typical behavior
- Intrusion Detection - monitoring systems and networks for unusual behavior
- Ecosystem Disturbances - try to predict events like hurricanes and floods
- Public Health - use medical statistic reports for diagnosis
- Medicine - use unusual symptoms or test result to indicate potential health problems

Challenges

- How many outliers are there in the data?
- Method is unsupervised
- There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data.

Anomaly Detection Schemes

General Steps

- Build a profile of the "normal" behavior, profile can be patterns or summary statistics for the overall population.
- Use the "normal" profile to detect anomalies, anomalies are observations whose characteristics differ significantly from the normal profile

Types of anomaly detection schemes

i. **Graphical based** : Box plot (1-D), Scatter plot (2-D), Spin plot (3-D)

ii. Statistical-based :

- Assume a parametric model describing the distribution of the data (e.g., normal distribution.
- A statistical test that depends on:
  . Data distribution
  . Parameter of distribution (e.g., mean, variance)
  . Number of expected outliers (confidence limit)

a. Grubbs' Test :

- Detect outliers in univariate data.
- Assume data comes from normal distribution.
- Detects one outlier at a time, remove the outlier, and repeat.

b. Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
  – M (majority distribution)
  – A (anomalous distribution)

General Approach:

– Initially, assume all the data points belong to M
– Let Lt(D) be the log likelihood of D at time t
- Let Lt+1 (D) be the new log likelihood.
- Compute the difference, $\Delta = Lt(D) – Lt+1 (D)$
- If $\Delta > c$ (some threshold), then Xt is declared as an anomaly and moved permanently from M to A

Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution

iii. Distance-based: Data is represented as a vector of features.

Three major approaches
– Nearest-neighbor based
– Density based
– Clustering based

iv. Model-based :

- An anomaly detection model predicts whether a data point is typical for a given distribution or not.
- An atypical data point can be either an outlier or an example of a previously unseen class.

- Normally, a classification model must be trained on data that includes both examples and counter-examples for each class so that the model can learn to distinguish between them.
- For example, a model that predicts side effects of a medication should be trained on data that includes a wide range of responses to the medication.

v. Convex Hull Method

- Extreme points are assumed to be outliers. Use convex hull method to detect extreme values.
- Major limitation is if the outlier occurs in the middle of the data.

Issues

a. **Number of Attributes**: Since an object may have many attributes, it may have anomalous values for some attributes; an object may be anomalous even if none of its attribute values are individually anomalous.

b. **Global Vs Local Perspective**: An object may seem unusual with respect to all objects, but not with respect to its local neighbors.

c. **Degree of Anomaly**: Some objects are more extreme anomalies than others;

d. **One at Time Vs Many at Once**: Is it better to remove anomalous objects one at a time or identify a collection of objects together?

e. **Evaluation**: Finding a good measure of evaluation for the process of anomaly detection when class labels are available and when class labels are not available.

f. **Efficiency**: calculate the computational cost of the process of anomaly detection scheme.

Base Rate Fallacy

- The base-rate fallacy is people's tendency to ignore base rates in favor of individuating information when such is available rather than integrate the two. This tendency has important implications for understanding judgment phenomena in many clinical, legal, and social-psychological settings.
- Base rate fallacy, also called base rate neglect or base rate bias, is a formal fallacy. If presented with related base rate information and specific information, the mind tends to ignore the former and focus on the latter.

Example

A group of policemen have breathalyzers displaying false drunkenness in 5% of the cases in which the driver is sober. However, the breathalyzers never fail to detect a truly drunk person. 1/1000 of drivers are driving drunk. Suppose the policemen then stop a driver at random, and force the driver to take a breathalyzer test. It indicates that the driver is

drunk. We assume you don't know anything else about him or her. How high is the probability he or she really is drunk?

Many would answer as high as 0.95, but the correct probability is about 0.02.

To find the correct answer, one should use Bayes' theorem. The goal is to find the probability that the driver is drunk given that the breathalyzer indicated he/she is drunk, which can be represented as

$$p(drunk|D)$$

where "D" means that the breathalyzer indicates that the driver is drunk.

Using Bayes' Theorem ,

$$p(drunk|D) = \frac{p(D|drunk)\,p(drunk)}{p(D)}$$

We have,

$$p(drunk) = 0.001$$
$$p(sober) = 0.999$$
$$p(D|drunk) = 1.00$$
$$p(D|sober) = 0.05$$

$$p(D) = p(D|drunk)\,p(drunk) + p(D|sober)\,p(sober)$$
$$p(D) = 0.05095$$

Putting values into Bayes' Theorem, we get

$$p(drunk|D) = 0.019627\cdot$$

A more intuitive explanation: in average, for every 1000 drivers tested,

- 1 driver is drunk, and it is 100% certain that for that driver there is a true positive test result, so there is 1 true positive test result
- 999 drivers are not drunk, and among those drivers there are 5% false positive test results, so there are 49.95 false positive test results therefore the probability that one of the drivers among the 1 + 49.95 = 50.95 positive test results really is drunk is $p(drunk|D) = 1/50.95 \approx 0.019627$. The validity of this result does, however, hinge on the validity of the initial assumption that the policemen stopped the driver truly at random, and not because of bad driving. If that or another non-arbitrary reason for stopping the driver was present, then the calculation also involves the probability of a drunk driver driving competently and a non-drunk driver driving competently.

Chapter- 7 Advanced Application

A. Web Mining

Web mining is the application of data mining techniques to extract knowledge from Web data, i.e. Web Content, Web Structure and Web Usage data.

Web Mining Taxonomy

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined.

a. Web Content Mining:

- Web Content Mining is the process of extracting useful information from the contents of Web documents.
- Content data corresponds to the collection of facts a Web page was designed to convey to the users.
- May consist of text, images, audio, video, or structured records such as lists and tables.
- Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages.

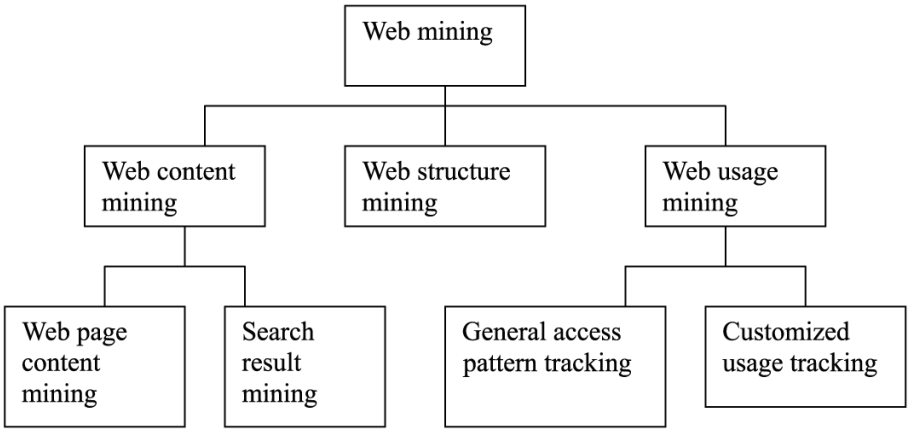
b. Web Structure Mining:

- The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages.
- Web Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used.
    - Hyperlinks: A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink.
    - Document Structure: In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model structures out of documents.

c. Web Usage Mining:

- Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications.

- Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.
- Web usage mining itself can be classified further depending on the kind of usage data considered:
    - Web Server Data: The user logs are collected by Web server. Typical data includes IP address, page reference and access time.
    - Application Server Data: Commercial application servers such as Web logic Story Server have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
    - Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these specially defined events. It must be noted however that many end applications require a combination of one or more of the techniques applied in the above the categories.



Challenges:

i. Too huge for effective data warehousing and data mining.
ii. Too complex and heterogeneous.
iii. Growing and changing rapidly
iv. Broad diversity of user communities.
v. Only small portion of the information on the web is truly relevant or useful.

The Page Rank Algorithm

The original Page Rank algorithm was described by Lawrence Page and Sergey Brin in several publications. It is given by

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

where
PR(A) is the Page Rank of page A,
PR(Ti) is the Page Rank of pages Ti which link to page A,
C(Ti) is the number of outbound links on page Ti and
d is a damping factor which can be set between 0 and 1.

- Page Rank does not rank web sites as a whole, but is determined for each page individually. Further, the Page Rank of page A is recursively defined by the Page Ranks of those pages which link to page A.
- The Page Rank of pages Ti which link to page A does not influence the PageRank of page A uniformly. Within the Page Rank algorithm, the Page Rank of a page T is always weighted by the number of outbound links C(T) on page T. This means that the more outbound links a page T has, the less will page A benefit from a link to it on page T.
- The weighted Page Rank of pages Ti is then added up. The outcome of this is that an additional inbound link for page A will always increase page A's Page Rank.
- Finally, the sum of the weighted Page Ranks of all pages Ti is multiplied with a damping factor d which can be set between 0 and 1. Thereby, the extend of PageRank benefit for a page by another page linking to it is reduced.

A Different Notation of the PageRank Algorithm

Lawrence Page and Sergey Brin have published two different versions of their Page Rank algorithm in different papers. In the second version of the algorithm, the Page Rank of page A is given as

$$PR(A) = (1-d) / N + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Where N is the total number of all pages on the web. The second version of the algorithm, indeed, does not differ fundamentally from the first one.

The Characteristics of Page Rank

The characteristics of Page Rank shall be illustrated by a small example.

We regard a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A. According to Page and Brin, the damping factor d is usually set to 0.85, but to keep the calculation simple we set it to 0.5. The exact value of the damping factor d admittedly has effects on Page Rank, but it does not influence the fundamental principles of Page Rank. So, we get the following equations for the Page Rank calculation:
$PR(A) = 0.5 + 0.5\ PR(C)$

PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

These equations can easily be solved. We get the following Page Rank values for the single pages:

PR(A) = 14/13 = 1.07692308
PR(B) = 10/13 = 0.76923077
PR(C) = 15/13 = 1.15384615

It is obvious that the sum of all pages' Page Ranks is 3 and thus equals the total number of web pages. As shown above this is not a specific result for our simple example. For our simple three-page example it is easy to solve the according equation system to determine Page Rank values. In practice, the web consists of billions of documents and it is not possible to find a solution by inspection.

The Iterative Computation of Page Rank

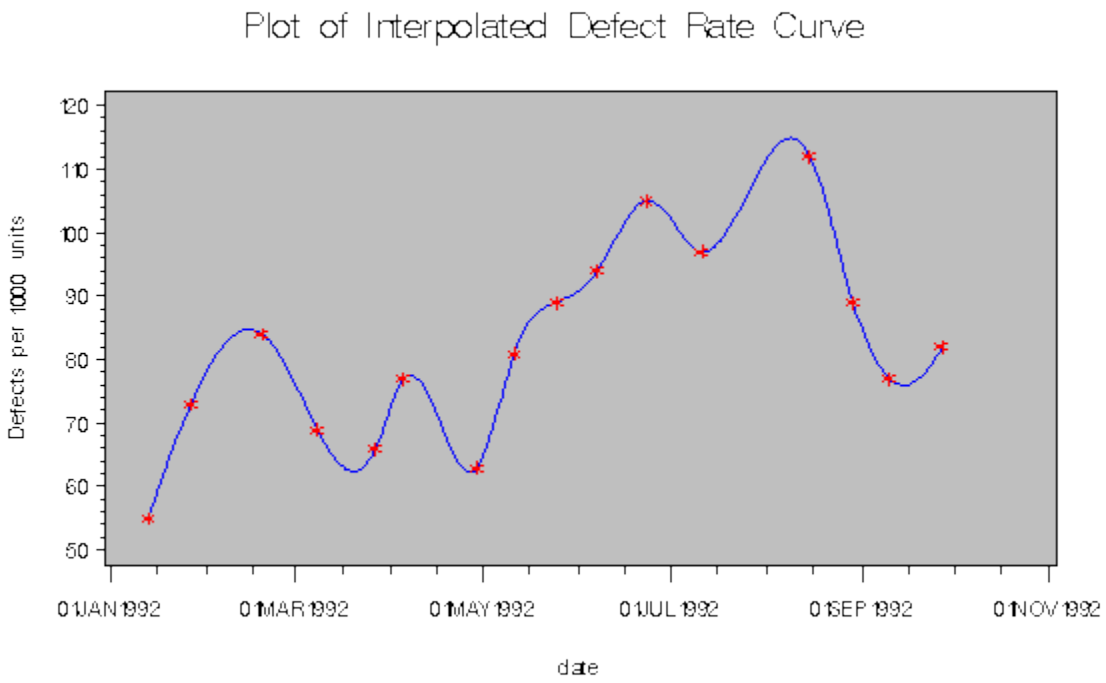
Because of the size of the actual web, the Google search engine uses an approximate, iterative computation of Page Rank values. Each page is assigned an initial starting value and the Page Ranks of all pages are then calculated in several computation circles based on the equations determined by the Page Rank algorithm. The iterative calculation shall again be illustrated by our three-page example, whereby each page is assigned a starting Page Rank value of 1.

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615

We see that we get a good approximation of the real Page Rank values after only a few iterations.

B. Time Series Data Mining

- Consists of sequences of values or events obtained over repeated measurement of time at equal time interval in most of the time.
- Used in application such as stock prediction, economic analysis etc.
- In general, there are two goals in time series analysis.
    i. Modeling Time Series: Generating the time series with underlying mechanism.
    ii. Forecasting Time Series: Predict the future values of the time series variables.



Major components for trend analysis in time series data

i. **Trend or Long term Movements**: Indicates the general direction in which a time series is moving over long or short interval of time through trend curve or trend line.
ii. **Cyclic Movement or Cyclic Variations**: Long term oscillations about a trend curve or line which may or may not be periodic.
iii. **Seasonal Movements or Variations**: These are systematic or calendar related. Eg. Sudden rise in sales of sweets in Tihar.
iv. **Irregular or Random Movements:** Series due to random or chance events. Eg. Price rise in crisis of supply.

Approaches for time series data analysis:

- Regression analysis is commonly used for find trend in time series data.
- Seasonal Index is used for analysis to adjust the reative values of a variable during the time series.
- Autocorrelation analysis is applied between i^{ith} element of the series and the $(i-k)^{th}$ element to detect seasonal patterns. Where K is referred to as the log.
- Calculating the moving average of order n is the common method for determining trend.
  Eg:
  Original Data: 3 7 2 0 4 5 9 7 2
  Moving average of order3: (3 + 7 + 2)/3 = 4, 3 2 3 6 7 6
  Weighted (1, 4, 1) average: ((1*3 +4*7 +1*2)/(1+4 +1))= 5.5, 2.5 1 3.5 5.5 8 6.5
- Free hand method is used to draw approximate curve or line to fit a set of data based on user's judgment.
- Least square method is used to fit best curve.

C. Object/ Image/ Multimedia Mining:

- Multimedia database system stores and manages a large collection of multimedia data such as audio, video, images, graphics, speech, text etc.
- Image/multimedia mining deals with extraction of implicit knowledge, data relationship or other patterns not explicitly stored in images/multimedia
- The fundamental challenges in images mining is to determine the low-level pixel representation contained in an image or image sequence and cane be effectively and efficiently processed to identify high level spatial objects and relationships.
- Typical image/multimedia processing involves preprocessing, transformations and feature extraction mining, evaluation and interpretation of the knowledge.
- Different data mining techniques can be used such as association rules, clustering.