Uniwersytet Ekonomiczny we Wrocławiu

Mining of Financial Databases

3. Clustering - Outlier Analysis

Jerzy KORCZAK
email: jerzy.korczak@ue.wroc.pl
http://www.korczak-leliwa.pl
http://citi-lab.pl

Outlier – simple understanding

An outlier is an extremely high or extremely low value in a data set.

How to find outliers?

Example

Number of customers

10, 12, 11, 15, 11, 14, 13, 17, 12, 22, 14, 11

Min ??? Max ???

Outlier – simple understanding

Simple method using IQR (InterQuartile Range J.Tukey)

Outlier is any value outside the range

[*Q1 – k (IQR), Q3 + k (IQR)*]

where *IQR= Q3 – Q1,* and *k=1.5*

Number of customers sorted

10, 11, 11, 11, 12, 12, 13, 14, 14, 15, 17, 22

Q1 = 11 *Q3=14.5* *IQR = 3.5*

*11 - 1.5 * 3.5 = 5.75*

*14 + 1.5 * 3.5 = 19.75*

So 10 > 5.75 is not an outlier; but 22 > 19.75 is an outlier.

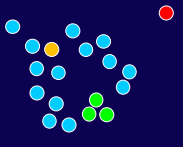
Definition

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" [Hawkins, 1980]

Outliers are also referred to as ***abnormalities, discordants, deviants, or anomalies***

Type of outliers:

- Global
- Contextual
- Collective

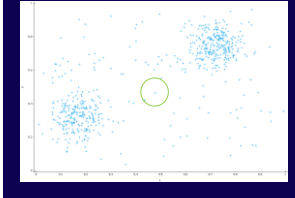


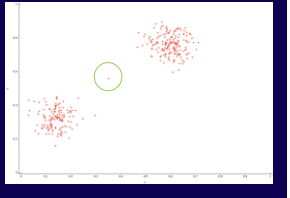
Basic Outlier Models

- Concepts – noise vs. anomaly
- Extreme Value Analysis
- Probabilistic and Statistical Models
- Linear Models
- Proximity-based Models
- Information Theoretic Models
- High-Dimensional Outlier Detection
- Meta-Algorithms for Outlier Analysis
- Financial applications
- Conclusions

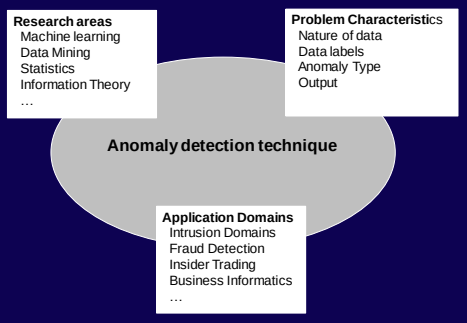
Concepts

The difference between noise and anomalies





Key components of anomaly detection technique



Z-value

- *Z-value test* for outlier analysis.

  Consider a set of 1-dimensional quantitative data observations, denoted by $X_1 \ldots X_N$, with mean μ and standard deviation σ. The Z value for the data point X_i is denoted by Z_i, and is defined as follows:

$$Z_i = \frac{|Xi - \mu|}{\sigma}$$

  A good "rule of thumb" is to use $Z_i \geq 3$ as a proxy for the anomaly

Anomaly testing

An instance is declared to be anomalous 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 N is the data size and $t^2_{\alpha/(2N),N-2}$ is a threshold used to declare an instance to be anomalous or normal. This threshold is the value taken by a t-distribution at a signifcance level of $\alpha/2N$.

Basic Outlier Models

- Several factor influence the choice of an outlier model, including the data type, data size, availability of relevant outlier examples, and the need for interpretability in a model.
- Levels of interpretability - intensional knowledge

Extreme Value Analysis

- The key is to determine the *statistical tails of the underlying distribution*.
- Problem:

  {1, 2, 2, 50, 98, 98, 99}

  is the outlier or not?

Probabilistic and Statistical Models

- The data is modeled in the form of a closed form probability distribution, and the parameters of this model are learned.
- The key assumption - the choice of the data distribution
- Example: a gaussian mixture model is a generative model; the parameters of these gaussian distributions are learned with the use of an *Expectation-Maximization (EM)* algorithm
- A key output of this method is the membership probability of the data points to the different clusters

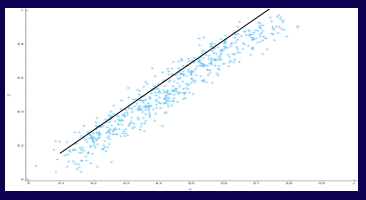
Probabilistic and Statistical Models (2)

- Detection of statistically *extreme univariate values*

  Examples:
  - In probabilistic modeling, the likelihood fit of a data point to the model is the outlier score.
  - In proximity-based modeling, the *k*-nearest neighbor distance, distance to closest cluster centroids, or local density value is the outlier score.
  - In linear modeling, the residual distance of a data point to a lower-dimensional representation of the data is the outlier score.
  - In temporal modeling, a function of the distance from previous data points (or the deviation from a forecasted value) is used to create the outlier score

Linear Models

- These methods model the data into lower dimensional embedded subspaces with the use of linear correlations

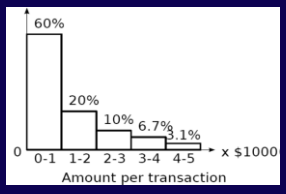


$$y_i = a \cdot x_i + b + \varepsilon_i \; \forall i \in \{1 \ldots N\}$$

Data points, which have large residuals (ε), are more likely to be outliers

Detection using histogram

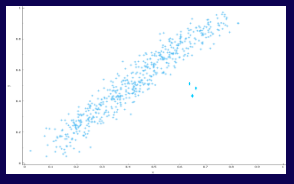
- The model of normal data is learned from the input data without any *a priori* structure.



- 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

Proximity-based Models

- The idea in proximity-based methods is to model outliers as points which are isolated from the remaining data.
- Methods: cluster analysis, density-based analysis or nearest neighbor analysis.



k-nearest neighbor

effective if $k \geq 3$

computationally expensive

Information Theoretic Models

- The key idea is to construct a *code book* in which to represent the data, and outliers are defined as points which removal results in the *largest decrease* in description length, or the most accurate summary representation in the same description length after removal

High-Dimensional Outlier Detection

- Challenge:
  - high dimensionality (to many features)
  - sparse data
  - data points become equidistant from one another
- Approaches:
  - Extending conventional outlier detection
  - Finding outliers in subspaces (lower D)
  - Modeling HD outliers (new heuristics)

Meta-Algorithms for Outlier Analysis

- ***Sequential ensembles***: a given algorithm or set of algorithms are applied sequentially, so that future applications of the algorithms are impacted by previous applications, in terms of either modifications of the base data for analysis or in terms of the specific choices of the algorithms. The final result is either a weighted combination of, or the final result of the last application of an outlier analysis algorithm.
- ***Independent ensembles***: different algorithms, or different instantiations of the same algorithm are applied to either the complete data or portions of the data. The choices made about the data and algorithms applied are independent of the results obtained from these different algorithmic executions. The results from the different algorithm executions are combined together in order to obtain more robust outliers.

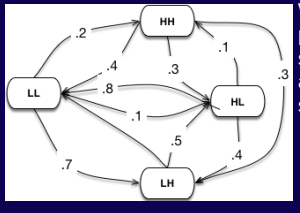
Financial Applications - Credit Card Fraud

A credit card company maintains the data corresponding to the card transactions by the different users. Each transaction corresponds to a set of attributes corresponding to the user identifier, amount spent, geographical location etc. The card company may also have labeled data containing previous examples of fraudulent transactions. It is desirable to determine fraudulent transactions from the data.

- The most common technique is to build **user profiles** on short segments of transaction sequences.
- The key is to design a similarity function
- The major challenge with anomaly detection in credit card data, is that **false positives** are extremely common, and **false negatives** are expensive,even when rare.

Univariate Collective Outliers – Markov chains

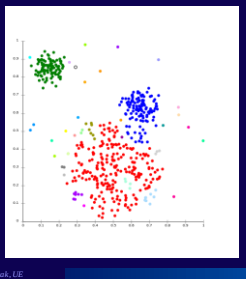
Credit card transactions using Markov chains: let's represent each transaction using two values: transaction value (L, H) and time since the last transaction (L, H). The states LL, LH, HL, HH and each transaction would be a transition from one state to another state.



We can find the probability of any new sequence happening and then mark rare sequences as **anomalies**.

Multivariate Collective Outliers – Clustering

Clustering the data, normal data will belong to clusters while anomalies will not belong to any clusters or belong to small clusters.



Improvement: manually inspect ranges of each cluster and label each cluster as anomalous or normal and use that while doing anomaly check.

Another approach: nearest neighbour technique

Financial Applications - Insurance Claim Fraud

- *In this case, claims are made by different entities on the basis of insurance policies. Significant anomalies need to be discovered from the data on this basis.*
- The problem is essentially an application of multidimensional (point) anomaly detection
- The key step is to extract the correct features from the insurance claim documents -> domain specific way
- Issues:
  - user-specific profiles cannot be constructed
  - repeated claims by a single user is often an indicator of fraud

Financial Applications - Stock Market Anomalies

- *The financial tickers of the different stocks and options correspond to time-series data streams. In some cases, significant anomalies may be created by external events. The early detection of such events may be useful in the determination of unknown influencing factors such as insider trading, or automated stock trading glitches*
- Ex. 2010 Flash Crash

  the use of spoofing, lyering, and front running are now prohibited

- Both deviation-based contextual point anomalies and time-series shape-based collective anomalies provide insights about the unusual interactions.

Conclusions

Key problems in data mining

Software resources available (KDD Nuggets website, SAS, SPSS, Weka, ELKI,…)

Numerous challenges:

- Difficulty in distnguishing between noise and anomalies
- Incorporation of human feedback
- Domain knowledge and explicit supervision