ANOMALY AND NOVELTY DETECTION

Michal Rozenwald

April 18, 2017

Faculty of Computer Science
National Research University Higher School of Economics

OUTLINES

INTRODUCTION

What are anomalies and novelty detection?

Challenges

APPLICATIONS

Techniques

I. Probabilistic Statistical

II. Distance-based

Nearest Neighbor

Clustering

III. Domain-based

Classification

IV. Reconstruction-based

Spectral Techniques

Neural Networks - AutoEncoder

V. Information-theoretic

Conclusion

References

INTRODUCTION

WHAT ARE ANOMALIES AND NOVELTY DETECTION?

Task: to decide whether a new observation belongs to the same distribution as existing observations (inlier), or should be considered as different (outlier)

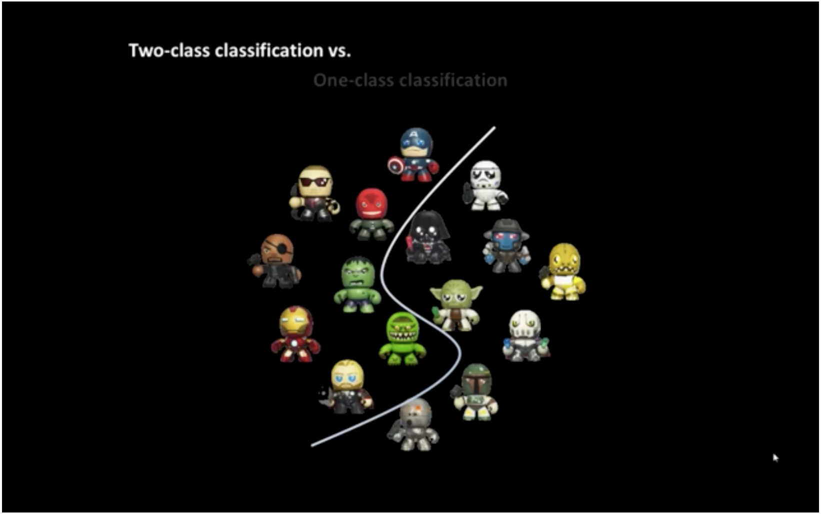
Anomaly or outlier - pattern in the data that does not conform to the expected behavior, objects that are different than most others

Anomaly/outlier detection:
The training data contains outliers, and we need to fit the central mode of the training data, ignoring the deviant observations.

Novelty - new and not resembling something formerly known or used

Novelty detection:
The training data is not polluted by outliers, and we are interested in detecting anomalies in new observations.

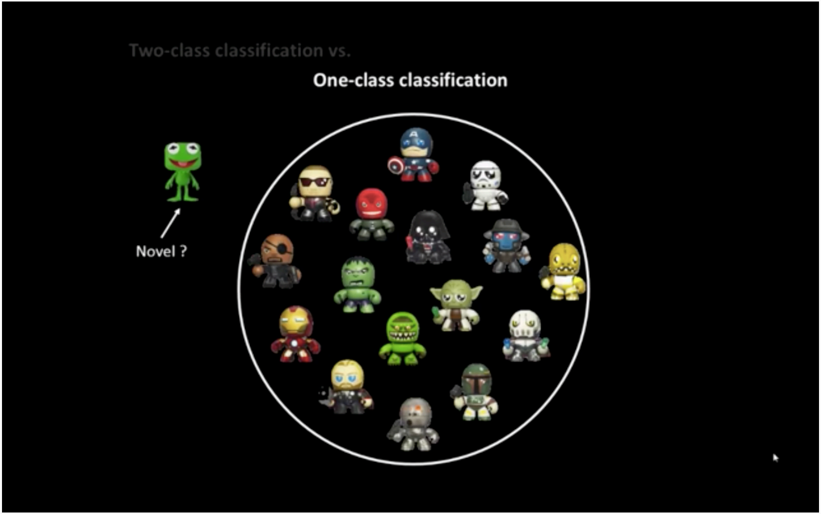
2-CLASS CLUSSIFICATION



Two-class classification vs.

One-class classification

1-CLASS CLUSSIFICATION



Two-class classification vs.

One-class classification

NOVELTY DETECTION AS ONE-CLASS CLASSIFICATION

Two-class classification problem:

$X = \{(x_i, y_i) | x_i \in \mathbb{R}^D, i = 1...N\}$ $x_i \in \mathbb{R}^D$ - input, $y_i \in \{-1, 1\}$ - label

$a(x) : \mathbb{R}^D \rightarrow [-1, 1]$ - estimate of one of the two labels is obtained $y = a(x'|X)$ for a given input vector x'

*General multi-class classification problems are often decomposed into multiple two-class classification problems

Novelty detection approach:

“normal” patterns X - training, “abnormal” - relatively few.

$M(\theta)$ - model of normality, θ - free parameters of the model

$z(x)$ - novelty scores for previously unseen test data x

Larger $z(x)$ - increased “abnormality” with respect to the model of normality.

$k := z(x)$ - novelty threshold - decision boundary: $z(x) \leq k$: x - “normal”, otherwise “abnormal”

CHALLENGES

Challenges

- # Outliers unknown
- Method is unsupervised
- Difficult to validate
- # Normal > # Abnormal

Found It!!!
Congratulations, it only took you 65298 seconds
www.jolyon.co.uk

- Boundary between normal and outlying behavior - not precise
- Insufficient data to describe “abnormalities”
- Many possible “abnormals”
- Defining a representative normal region is hard
- Exact notion of anoutlier is different for different applications
- Data might contain noise
- Normal behavior keeps evolving

DATASETS PROBLEM

Typical datasets:

- very large number of examples of the “normal” condition (also known as positive examples) is available

- insufficient data to describe “abnormalities” (also known as negative examples)

Possible solution: Manipulating Data Records

MANIPULATING DATA RECORDS

- **Over-sampling the rare class**

– Make the duplicates of the rare events until the data set contains as many examples as the majority class => balance the classes

– Does not increase information but increase misclassification cost

- **Down-sizing (undersampling) the majority class**

– Sample the data records from majority class

– Introduce sampled data records into the original data set instead of original data records from the majority class

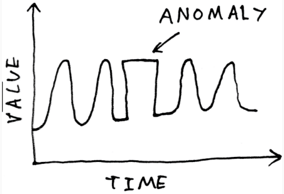
– Usually results in a general loss of information

- **Generating artificial anomalies**

– SMOTE (Synthetic Minority Over-sampling TEchnique) - new rare class examples are generated inside the regions of existing rare class examples

ASPECTS OF ANOMALY DETECTION PROBLEM

- Nature of Input Data:
Binary, Categorical, Continuous, Hybrid
- Relationship among data instances
- Availability of supervision
- Type of Anomaly:
point, contextual, structural
- Output of anomaly detection
- Evaluation of anomaly detection techniques



DATA LABELS: AVAILABILITY OF SUPERVISION

- **Supervised Anomaly Detection**

– Labels available for both normal data and anomalies

– Similar to rare class mining

- **Semi-supervised Anomaly Detection**

– Labels available only for normal data

- **Unsupervised Anomaly Detection**

– No labels assumed

– Based on the assumption that anomalies are very rare compared to normal data

EVALUATION AND SCORING

Effectiveness of ND techniques:
- how many novel data points are correctly identified
- how many normal data are incorrectly classified as novel data (false alarm rate)

Receiver operating characteristic **(ROC)** curves - trade-off between detection rate and false alarm rate

Aim: high detection rate & low false alarm rate

Efficiency of ND: - computational cost: time and space complexity - amount of memory required to implementation

APPLICATIONS

БОЛЬШАЯ АКТУАЛЬНОСТЬ ПРОБЛЕМАТИКИ

+ Автоматическое обнаружение аномального человеческого поведения при видеонаблюдении [Pham и др. 2014]
+ Мониторинг и анализ смертности и заболеваемости раком легких [Dass 2009; Dass и др. 2011; Taweab и др. 2015]
+ Мониторинг уровня хлора в питьевой воде [Gu ́epi ́e и др. 2012]
+ Обнаружение изменений структуры породы при бурении скважин [Adams и др. 2007]
+ Диагностика задержки внутриутробного роста [Petzold и др. 2004]
+ Мониторинг целостности системы геопозиционирования [M Basseville и др. 2002]
+ Обнаружение возникновения эпидемий [MacNeill и др. 1995]
+ Обнаружение и изоляция отказов узлов систем управления транспортными средствами[Malladi и др. 1999; Willsky 1976]
+ Обнаружение аритмии (внезапных изменений ритма биения) сердца [Willsky 1976]

АКТУАЛЬНОСТЬ ПРОБЛЕМАТИКИ. ОБЪЕМ ПУБЛИКУЕМОЙ ЛИТЕРАТУРЫ

Поиск в системе индексации **Google Scholar** выдает, начиная с 2000 года:

* change point detection — 10 200 статей

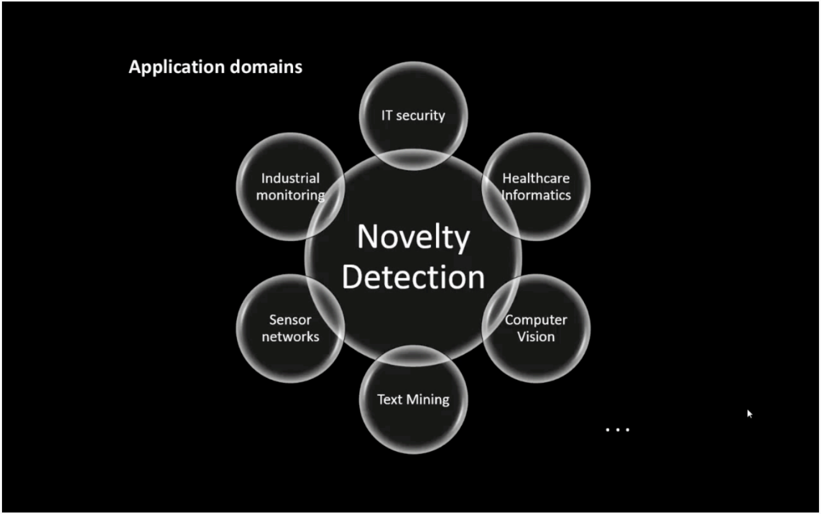
* anomaly detection — 53 900 статей

* break detection — 3 980 статей

* обнаружение разладок, обнаружение аномалий, обнаружение изменений — 765 статей

Первые работы по разладкам: 1931 год, W. A. Shewhart (цель — контроль качества выпускаемой продукции).
А. В. Артёмов Стохастические задачи о разладке

APPLICATIONS



APPLICATIONS OF ANOMALY DETECTION

- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining

APPLICATIONS: INTRUSION DETECTION

Intrusion Detection: –process of monitoring the events occurring in a computer system or network and analyzing them for intrusions

Intrusions are defined as attempts to bypass the security mechanisms of a computer or network

Challenges:

– Traditional signature-based intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats

– Substantial latency in deployment of newly created signatures across the computer system

Anomaly detection can alleviate these limitations

APPLICATIONS: MEDICAL AND PUBLIC HEALTH ANOMALY DETECTION

Detect anomalous patient records

– Indicate disease outbreaks, instrumentation errors, etc.

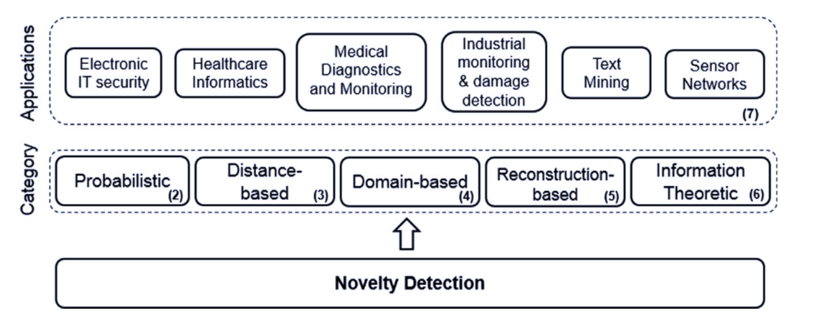
Key Challenges:

– Only normal labels available
– Misclassification cost is very high
– Data can be complex: spatio-temporal

TECHNIQUES

CLASSIFICATION OF NOVELTY DETECTION TECHNIQUES

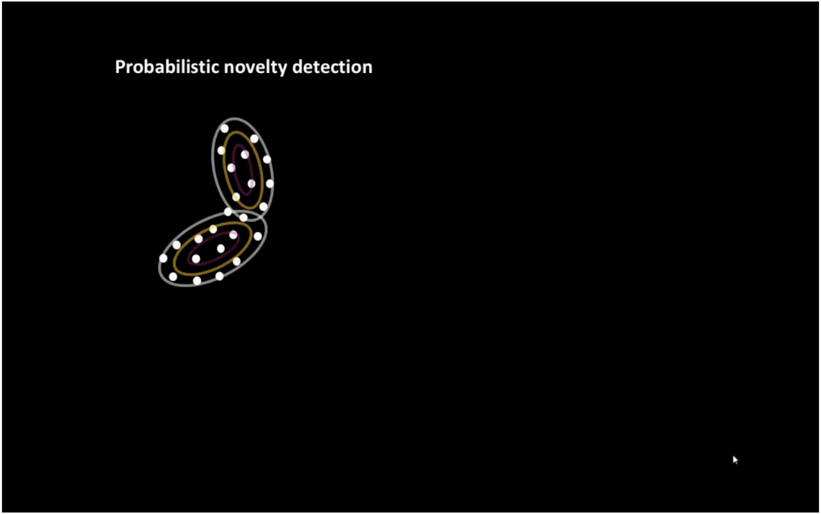
Categories of approaches:
(i) Probabilistic, Statistical
(ii) Distance-based
(iii) Domain- based
(iv) Reconstruction-based
(v) Information-theoretic



I. PROBABILISTIC NOVELTY DETECTION

Probabilistic novelty detection

I. PROBABILISTIC NOVELTY DETECTION



Probabilistic novelty detection

PROBABILISTIC NOVELTY DETECTION

Probabilistic methods:
often involve a density estimation of the “normal” class.

Assumption:
Low density areas in the training set indicate that these areas have a low probability of containing “normal” objects

STATISTICAL NOVELTY DETECTION

Key Assumption:
Normal data instances occur in high probability regions of a statistical distribution, while anomalies occur in the low probability regions of the statistical distribution

General Approach:
Estimate a statistical distribution using given data, and then apply a statistical inference test to determine if a test instance belongs to this distribution or not

Typical Test: If an observation is more than 3 standard deviations away from the sample mean, it is an anomaly

TYPES OF STATISTICAL TECHNIQUES

- **Parametric Techniques**

– Assume that the normal (and possibly anomalous) data is generated from an underlying parametric distribution

– Learn the parameters from the training sample

- **Non-parametric Techniques**

– Do not assume any knowledge of parameters

– Use non-parametric techniques to estimate the density of the distribution – e.g., histograms

EXAMPLE: Using Chi-square Statistic

Normal data is assumed to have a multivariate normal distribution
Sample mean is estimated from the normal sample
Anomaly score of a test instance is

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{\hat{X}_i}$$

PROBABILISTIC NOVELTY DETECTION

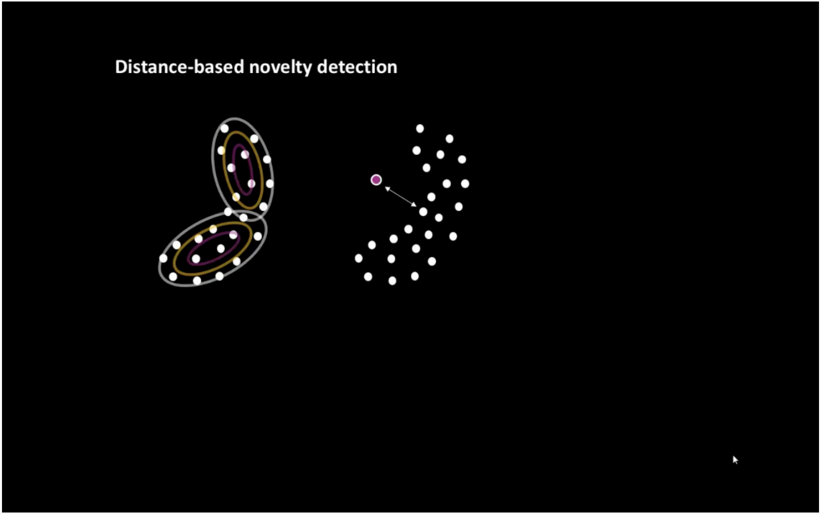
Evaluation: Performance is limited when the size of the training set is very small, particularly in moderately high-dimensional spaces.

As the dimensionality increases, the data points are spread through a larger volume in the data space.

In many real-life scenarios, no a priori knowledge of the data distributions is available, and so parametric approaches may be problematic if the data do not follow the assumed distribution.

Thus, non-parametric techniques are appealing since they make fewer assumptions about the distribution characteristics.

II. DISTANCE-BASED NOVELTY DETECTION



Distance-based novelty detection

DISTANCE-BASED NOVELTY DETECTION

Distance-based methods:
includes the concepts of **nearest-neighbour** and **clustering analysis** that have also been used in classification problems.

These methods rely on well-defined distance metrics to compute the distance (similarity measure) between two data points.

Assumption:
“normal” data are tightly clustered, while novel data occur far from their nearest neighbours

NEAREST NEIGHBOR BASED TECHNIQUES

Key assumption: normal points have close neighbors while anomalies are located far from other points

General two-step approach

1. Compute neighborhood for each data record
2. Analyze the neighborhood to determine whether data record is anomaly or not

Categories:

– Distance based methods
Anomalies are data points most distant from other points

– Density based methods
Anomalies are data points in low density regions instances in low density regions as potential anomalies

NEAREST NEIGHBOR BASED TECHNIQUES

Advantages:

– Can be used in unsupervised or semi-supervised setting
- Do not make any assumptions about data distribution

Drawbacks:

– If normal points do not have sufficient number of neighbors the techniques may fail
– Computationally expensive
– In high dimensional spaces, data is sparse and the concept of similarity may not be meaningful anymore. Due to the sparseness, distances between any two data records may become quite similar
=> Each data record may be considered as potential outlier!

CLUSTERING BASED ANOMALY DETECTION TECHNIQUES

Key Assumption:
Normal data instances belong to large and dense clusters, while anomalies do not belong to any significant cluster

General Approach:

– Cluster data into a finite number of clusters
– Analyze each data instance with respect to its closest cluster
– Anomalous Instances:

  · Data instances that do not fit into any cluster (residuals from clustering)

  · Data instances in small clusters

  · Data instances in low density clusters

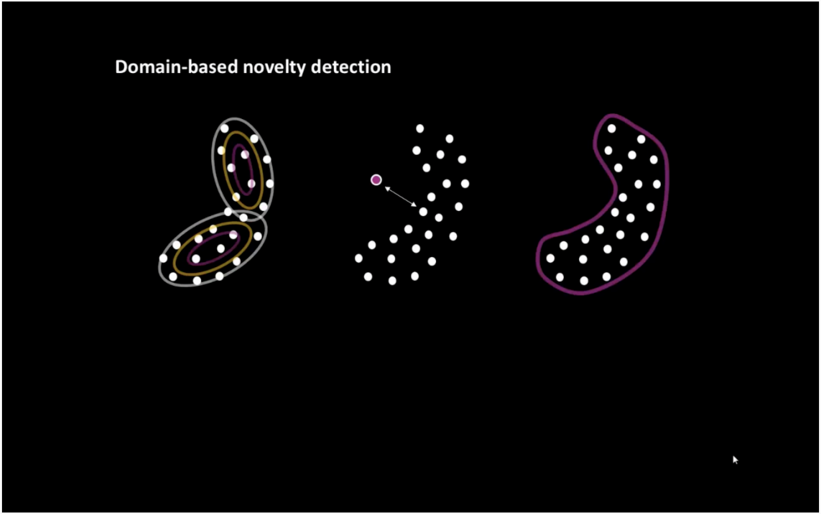
  · Data instances that are far from other points with in the same cluster

CLUSTERING BASED ANOMALY DETECTION TECHNIQUES

Advantages:
– Unsupervised
– Existing clustering algorithms can be plugged in

Drawbacks:
– If the data does not have a natural clustering or the clustering algorithm is not able to detect the natural clusters, the techniques may fail
– Computationally expensive
– In high dimensional spaces, data is sparse and distances between any two data records may become quite similar

III. DOMAIN-BASED NOVELTY DETECTION



Domain-based novelty detection

III. DOMAIN-BASED NOVELTY DETECTION

Domain-based methods:
try to describe a domain containing “normal” data by defining a boundary around the “normal” class such that it follows the distribution of the data, but does not explicitly provide a distribution in high-density regions.

Assumption:
Class membership of unknown data is determined by their **location** with respect to the boundary.

III. DOMAIN-BASED NOVELTY DETECTION

Method evaluation :
Domain-based approaches determine the location of the novelty boundary using only those data that lie closest to it and do not rely on the properties of the distribution of data in the training set.

Drawback: complexity associated with the computation of the kernel functions.
Overcome this problem: the choice of the appropriate kernel function may also be problematic.

*Not easy to select values for the parameters which control the size of the boundary region.

CLASSIFICATION BASED ANOMALY DETECTION TECHNIQUES

Main idea:
build a classification model for normal (and anomalous (rare)) events based on labeled training data, and use it to classify each new unseen event

Classification models must be able to handle skewed (imbalanced) class distributions

- **Supervised** classification techniques
  - Require knowledge of both normal and anomaly class
  - Build classifier to distinguish between normal and known anomalies
- **Semi-supervised** classification techniques
  - Require knowledge of normal class only!
  - Use modified classification model to learn the normal behavior and then detect any deviations from normal behavior as anomalous

CLASSIFICATION BASED ANOMALY DETECTION TECHNIQUES

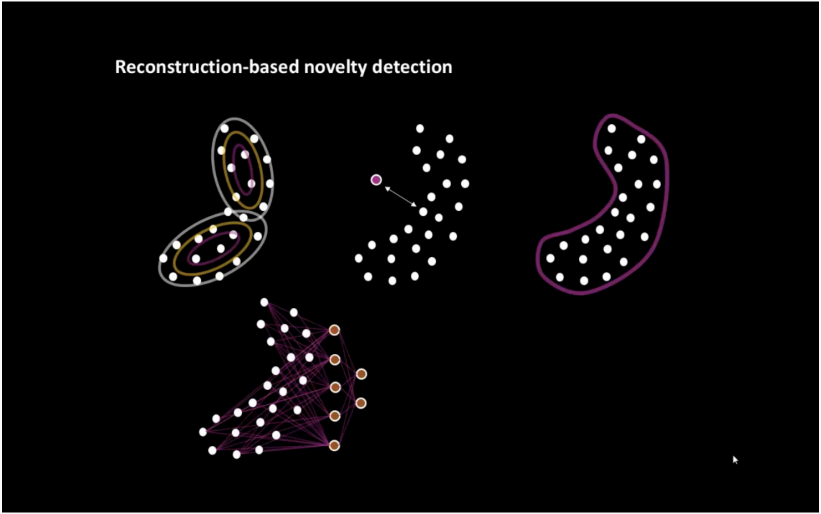
Advantages:

- **Supervised** classification techniques
  - Models that can be easily understood
  - High accuracy in detecting many kinds of known anomalies
- **Semi-supervised** classification techniques
  - Models that can be easily understood
  - Normal behavior can be accurately learned

Drawbacks:

- **Supervised** classification techniques
  - Require both labels from both normal and anomaly class
  - Cannot detect unknown and emerging anomalies
- **Semi-supervised** classification techniques
  - Require labels from normal class
  - Possible high false alarm rate - previously unseen (yet legitimate) data records may be recognized as anomalies

IV. RECONSTRUCTION-BASED NOVELTY DETECTION



Reconstruction-based novelty detection

IV. RECONSTRUCTION-BASED NOVELTY DETECTION

Reconstruction-based novelty detection:
involves training a regression model using the training set.

When "abnormal" data are mapped using the trained model, the reconstruction error between the regression target and the actual observed value gives rise to a high novelty score

They can autonomously model the underlying data, and when test data are presented to the system, the reconstruction error, defined to be the distance between the test vector and the output of the system, can be related to the novelty score

Neural networks: Multi-layer perceptron Hopfield networks, Autoassociative networks, Radial basis function, Self-organising networks

IV. RECONSTRUCTION-BASED NOVELTY DETECTION

Evaluation:

Networks require the optimisation of a predefined number of parameters that define the structure of the model, and their performance may be very sensitive to these model parameters.

Difficult to train in high-dimensional spaces.

Networks that use constructive algorithms (the structure of the model is allowed to grow) additional problem: select the most effective training method to enable the integration of new units into the existing model structure, and an appropriate stopping criterion (for when to stop adding new units).

Difficult to determine which are the key attributes and it is computationally expensive to estimate the correlation matrix of normal patterns accurately.

SPECTRAL TECHNIQUES

Analysis based on decomposition of data

Key Idea:

– Find combination of attributes that capture bulk of variability

– Reduced set of attributes can explain normal data well, but not necessarily the anomalies

Advantages:

– Can operate in an unsupervised mode

Drawback:

– Based on the assumption that anomalies and normal instances are distinguishable in the reduced space

AUTOENCODER (AE)

Auto Encoder - artificial neural network used for unsupervised learning of efficient codings.

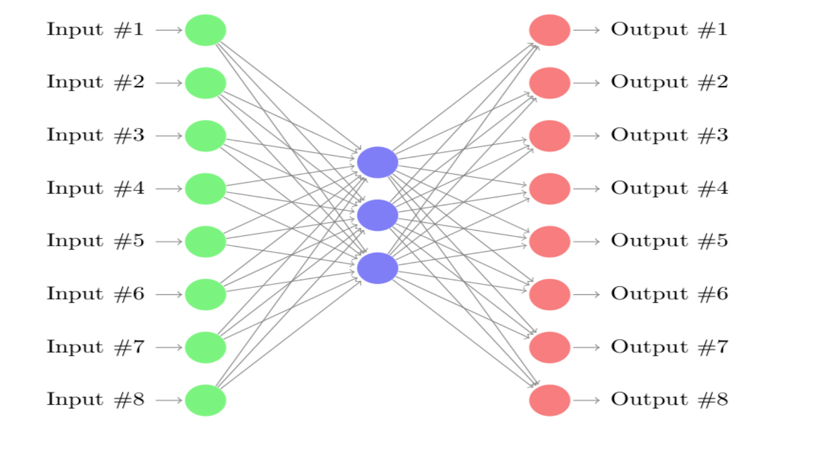
Goal: to induce a representation (encoding) for a set of data by learning an approximation of the identity function of this data Id : $\mathcal{X} \to \mathcal{X}$

In contrast to multilayer perceptrons (MLP):
number of output nodes = number of input nodes $\Rightarrow$ instead of being trained to predict target value Y given inputs X , autoencoders are trained to reconstruct their own inputs X (Unsupervised learning models)

AUTO ENCODER

Auto-encoder: 1 hidden layer, vectors length d=8

Attempt to reduce the dimensionality of the inputs to p=3



AUTO ENCODER

Autoencoder = ϕ -encoder + ψ-decoder

$\phi : \mathcal{X} \rightarrow \mathcal{F}, \psi : \mathcal{F} \rightarrow \mathcal{X}$

$\mathrm{argmin}_{\phi,\psi} ||X - (\phi \circ \psi)X||^2$

Simplest case - 1 hidden layer: $X \in \mathbb{R}^d$ maped to $z \in \mathbb{R}^p = \mathcal{F}$, (p<d):

$z = \sigma_1(Wx + b)$ - latent representation

$x' = \sigma_2(W'z + b')$ - reconstruction

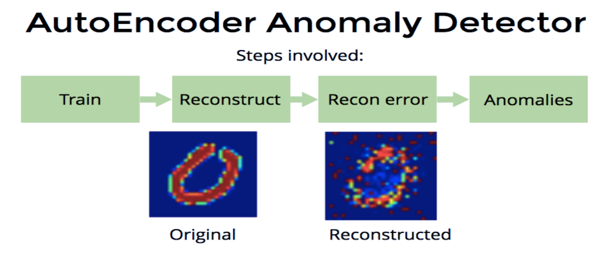
Train to minimise reconstruction errors (squared errors):

$\mathcal{L}(x, x') = ||x - x||^2 = ||x - \sigma_2(W'(\sigma_1(Wx + b)) + b')||^2 \rightarrow \min$

Autoencoding can also be seen as a non-linear alternative to PCA.

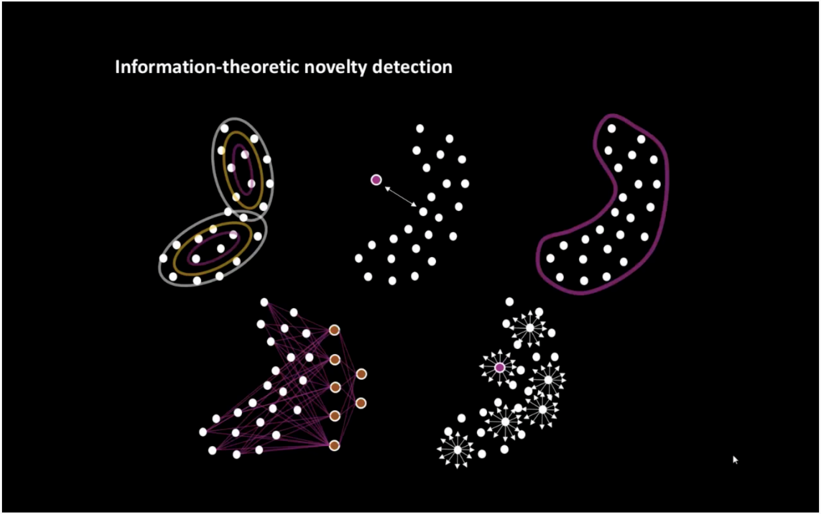
USING AUTO ENCODERS FOR ANOMALY DETECTION

Idea (Straightforward):



- Train an auto-encoder on $\mathcal{X}_{\text{train}}$ with good regularization
- Choose a threshold - like 2 standard deviations from the mean - which determines whether a value is an outlier (anomalies) or not. Threshold can be dynamic and depends on the previous errors (moving average, time component)

V. INFORMATION-THEORETIC NOVELTY DETECTION



Information-theoretic novelty detection

V. INFORMATION-THEORETIC NOVELTY DETECTION

Information-theoretic novelty detection methods:
compute the information content of a dataset using measures such as entropy, relative entropy, Kolmogorov complexity etc.

Assumption:
These methods assume that novelty significantly alters the information content of the otherwise "normal" dataset

Typically, metrics are calculated using the whole dataset and then that subset of points whose elimination from the dataset induces the biggest difference in the metric is found
This subset is then assumed to consist of novel data

V. INFORMATION-THEORETIC NOVELTY DETECTION

EXAMPLE:
local-search heuristic approach, which involves entropy analysis, to identify outliers

Entropy in information theory is a measure of the uncertainty associated with a random variable

Entropy function can be used to measure the degree of disorder of the remaining dataset after removal of high-entropy points

A point is considered to be an outlier if the entropy of the dataset decreases after its removal, compared with the entropy of the dataset after removal of all previous outlier candidates

This procedure is repeated until k outliers are identified

V. INFORMATION-THEORETIC NOVELTY DETECTION

Evaluation:

Information-theoretic approaches to novelty detection typically do not make any assumptions about the underlying distribution of the data

They require a measure that is sensitive enough to detect the effects of novel points in the dataset

Drawback:

techniques is very dependent on the choice of the information-theoretic measure

Measures can detect the presence of novel data points only if there is a significantly **large number of novel** data points present in the dataset

Computationally expensive, solution: approximations

*It may be difficult to associate a novelty score with a test point using an information theoretic-based method

VISUALIZATION BASED TECHNIQUES

Idea:
Anomalies are detected visually

Use visualization tools to observe the data
Provide alternate views of data for manual inspection

Advantages:
– Keeps a human in the loop.

Drawbacks:
– Works well for low dimensional data
– Anomalies might be not identifiable in the aggregated or partial views for high dimension data
– Not suitable for real-time anomaly detection

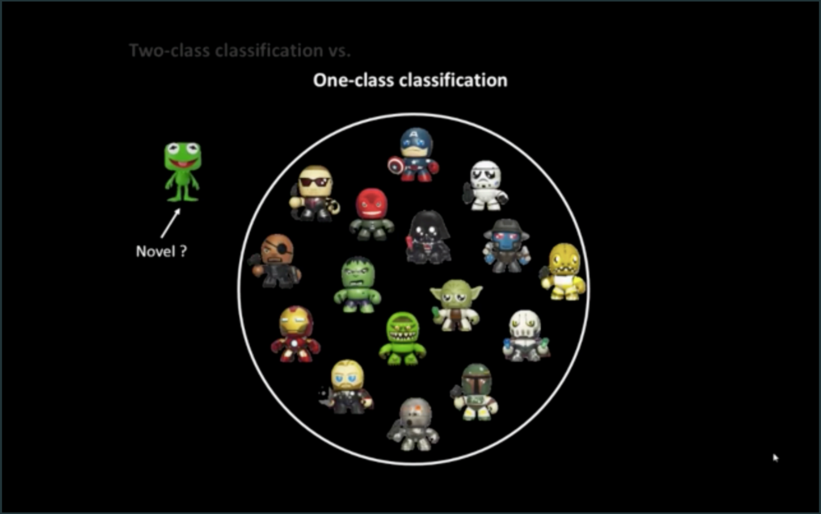
CONCLUSION

SUMMARY

- Novelty detection **main goal:** construct classifiers when only one class is well-sampled and well-characterised by the training data
- Need different approaches to solve a particular problem formulation. Is it possible to suggest what an **"optimal" method** of novelty detection would be
- **Complexity** (In General) :
 Probabilistic, reconstruction-based and domain-based methods have lengthy training phases, but with rapid testing (models can be trained offline)

Distance-based and information theoretic-based methods are computationally expensive in the test phase, which may be an important limitation in real-world settings

QUESTIONS?


Two-class classification vs.
One-class classification
Novel ?

REFERENCES

REFERENCES

- A review of novelty detection - Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. Signal Processing, 99: 215–249, 2014. (Cited on pages 3, 5, 7, and 29.) `https://www.youtube.com/watch?v=Lq_p34RvO2E`
- Anomaly Detection – A Survey, Varun Chandola, Arindam Banerjee, and Vipin Kumar, To Appear in ACM Computing Surveys 2008.
- Outlier detection techniques for wireless sensor networks: a survey - Y. Zhang, N. Meratnia, P. Havinga, IEEE Commun. Surv. Tutor. 12 (2) 2010 159–170.
- Data Mining for Anomaly Detection - Aleksandar Lazarevic United Technologies Research Center.
- Anomaly Detection in Time Series using Auto Encoders: http://philipperemy.github.io/anomaly-detection/
-

THANK YOU!