Data Mining II

Outlier Detection

Rita P. Ribeiro

2016/2017

Computer Science Department

U. PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Summary

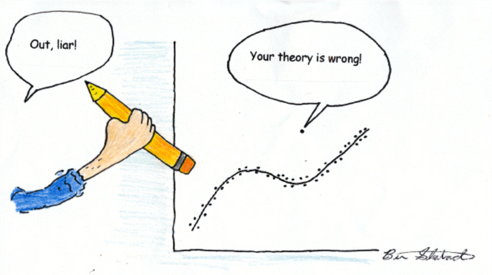
Outlier Detection
Basic Concepts

Motivation

- Most of data mining tasks focus on creating a model of the “normal” patterns in the data, extracting knowledge from what is common (e.g. frequent patterns).
- Still, rare patterns can also give us some import insights about data.
- Depending on the goal, those insights can be even more interesting/critical than the “normal” patterns.

What is an Outlier?

- *"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)



What is an Outlier? (cont.)

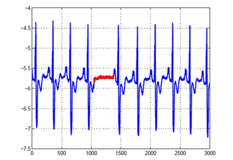
- Outliers represent patterns in data that do not conform to a well defined notion of normal.

- Initially, outliers were considered errors and their identification had data cleaning purposes.

- However, they can represent truthful deviation of data.
- For some applications, they represent critical information, which can trigger preventive or corrective actions.

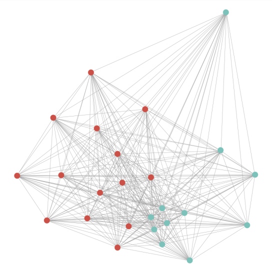
Where can Outliers occur?

Medical Analysis

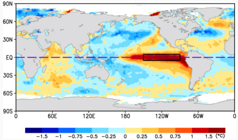


Financial Markets

Social Network Analysis

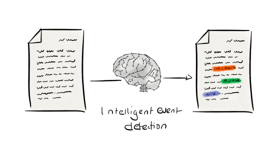


Anomalous Weather Patterns



Fraud Detection

Event Detection in Text/Social Media



Applications of Outlier Detection

- Quality Control and Fault Detection Applications
  - Quality Control
  - Fault Detection and Systems Diagnosis
  - Structure Defect Detection
- Financial Applications
  - Credit Card Fraud
  - Insurance Claim Fraud
  - Stock Market Anomalies
  - Financial Interaction Networks
- Intrusion and Security Applications
  - Host-based Intrusions
  - Network Intrusion Detection
- Web Log Analytics
  - Web Log Anomalies

Applications of Outlier Detection (cont.)

- Market Basket Analysis
  - Outlier transactions in association patterns
- Medical Applications
  - Medical Sensor Diagnostics
  - Medical Imaging Diagnostics
- Text and Social Media Applications
  - Event Detection in Text and Social Media
  - Spam Email
  - Noisy and Spam Links
  - Anomalous Activity in Social Networks
- Earth Science Applications
  - Sea Surface Temperature Anomalies
  - Land Cover Anomalies
  - Harmful Algae Blooms

Challenges of Outlier Detection

- Define every possible “normal” behaviour is hard.
- The boundary between normal and a outlying behaviour is often not precise.
- There is no general outlier definition; it depends on the application domain.
- It is difficult to distinguish real meaningful outliers from simple random noise in data.
- The outlier behaviour may evolve with time.
- Malicious actions adapt themselves to appear as normal.
- Inherent lack of known labeled outliers for training/validation of models.

Key Aspects of Outlier Detection Problem

- Nature of Input Data
- Type of Outliers
- Intended Output
- Learning Task
- Performance Metrics

Nature of Input Data

- Each data instance has:
  - One attribute (univariate)
  - Multiple attributes (multivariate)
- Relationship among data instances:
  - None
  - Sequential / Temporal
  - Spatial
  - Spatio-temporal
  - Graph
- Dimensionality of data

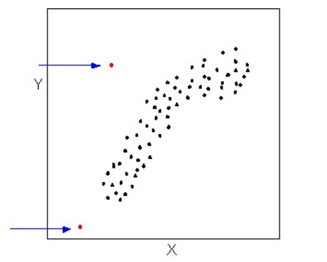
Types of Outliers

- Point (or Global) Outlier
- Contextual Outlier
- Collective Outlier

Types of Outliers (cont.)

Point Outlier

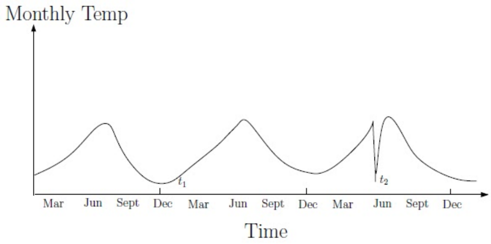
An instance that individually or in small groups is very different from the rest of the instances.



Types of Outliers (cont.)

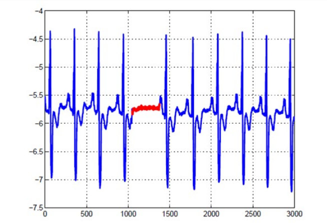
Contextual Outlier

An instance that when considered within a context is very different from the rest of the instances.



Collective Outlier

An instance that, even though individually may not be an outlier, inspected in conjunction with related instances and with respect to the entire data set is an outlier.



Intended Output

- Assign a label/value: identification normal or outlier instance.
- Assign a score: probability of being an outlier.
  - It allows the output to be ranked.
  - Requires the specification of a threshold.

Learning Task

Unsupervised Outlier Detection

- data set has no information on the behaviour of each instance;
- it assumes that instances with normal behaviour are far more frequent;
- most common case in real-life applications.

Semi-supervised Outlier Detection

- data set has a few instances of normal or outlier behaviour;
- some real-life applications, such as fault detection, provide such data.

Supervised Outlier Detection

- data set has instances of both normal and outlier behaviour;
- hard to obtain such data in real-life applications.

Performance Metrics

Inadequacy of Standard Performance Metrics

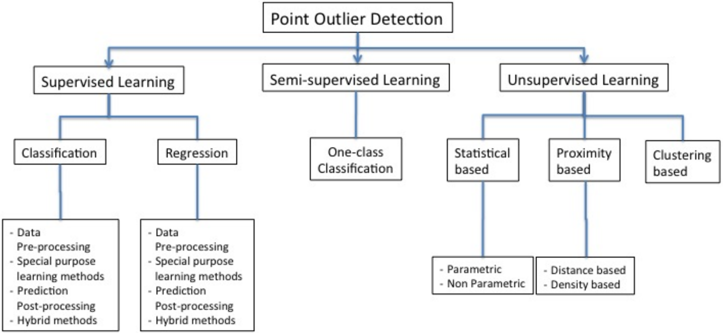
- Standard performance metrics (e.g. *accuracy*, *error rate*) assume that all instances are equally relevant for the model performance.
- These metrics would give a good performance estimation to a model that performs well on normal (frequent) cases and bad on outlier (rare) cases.

  ### Credit Card Fraud Detection:

  - data set D with only 1% of fraudulent transactions;
  - model M predicts all transactions as non-fraudulent;
  - M has a estimated accuracy of 99%;
  - yet, all the fraudulent transactions were missed!

Outlier Detection Approaches

Taxonomy of Outlier Detection Methods



Outlier Detection Approaches

Unsupervised Learning Techniques

Statistical-based Outlier Detection

Proposal

- All the points that satisfy a statistical discordance test for some statistical model are declared as outliers.

Advantages

- If the assumptions of the statistical model hold true, these techniques provide a justifiable solution for outlier detection.
- The outlier score is associated with a confidence interval.

Techniques

- Parametric
- Non-parametric

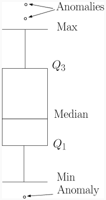
Assume one of the known probability distribution functions.

- *Grubbs' Test* (Grubbs, 1950)
  A statistical test used to detect outliers in a **univariate** data set assumed to come from a normally distributed population.

- Boxplot (Tukey, 1977)
  It assumes a near-normal distribution of the values in a **univariate** data set, and identifies as outlier any value outside the interval

$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

where Q_1 (Q_3) is the 1st (3rd) quartile and *IQR* is the interquartile range.



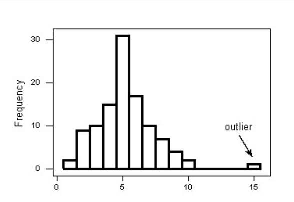
Statistical-based Outlier Detection: Parametric Techniques (cont.)

- *Mahalanobis* distance (Mahalanobis, 1936)
  - It assumes a multivariate normal distribution of data.
  - Incorporates dependencies between attributes by the covariance matrix.
  - Transforms a **multivariate** outlier detection task into a univariate outlier detection problem.
  - All the points with a large *Mahalanobis* distance are indicated as outliers.
- Mixture of parametric distributions
- etc.

Statistical-based Outlier Detection: Non-parametric Techniques

The probability distribution function is not assumed, but estimated from data.

- Histograms
  - Used for both univariate and multivariate data. For the later, the attribute-wise histograms are constructed and an aggregated score is obtained.
  - Hard to choose the appropriate bin size.



- Kernel functions
  - Adopt a kernel density estimation to estimate the probability density distribution of the data.
  - Ouliers are in regions of low density.

Statistical-based Outlier Detection

Disadvantages

- The data does not always follows a statistical model.
- Choosing the best hypothesis test statistics is not straightforward.
- Capture interactions between attributes is not always possible.
- Estimating the parameters for some statistical models is hard.

Proximity-based Outlier Detection

Proposal

- Normal instances occur in dense neighbourhoods, while outliers occur far from their closest neighbours.

Advantages

- Purely data driven technique
- Does not make any assumptions regarding the underlying distribution of data.

Some Techniques

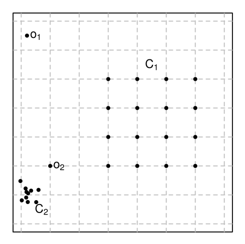
- Distance-based
- Density-based

Proximity-based Outlier Detection: Distance-based Techniques

A case *c* is an outlier if less than *k* cases are within a distance λ of *c* [Knorr and Ng, 1998]

- Outliers are points far away from other points, thus given a distance metric there should not be a lot of other points in their neighborhood.
- Define proper distance metric (e.g euclidean distance)
  - The notion of distance between cases with many variables may be distorted by different scales, different importance, different types (numerical, nominal)
- Define a “reasonable” neighborhood (λ)
- Define what is “a lot of other points” (*k*)

- Major cost: for each point is calculated its distance to all the other points.
- Optimization algorithms include index-based, cell-based approaches.
- The use of **global distance** measures poses difficulties in detecting outliers in data sets with different density regions.
- Example:



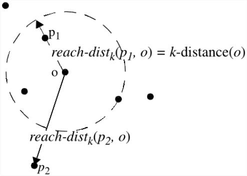
- o_1 and o_2 are outliers
- but, for the point o_2 to be identified as an outlier, all the points in C_1 would have to be identified as outliers too.

Proximity-based Outlier Detection: Density-based Techniques

- Concept of outliers should be **locally** inspected.
- Compare points to their local neighborhood, instead of the global data distribution
- The density around an outlier is significantly different from the density around its neighbours.
- Use the relative density of a point against its neighbours as the indicator of the degree of the point being an outlier.
- Outliers are points in lower local density areas with respect to the density of its local neighbourhood.

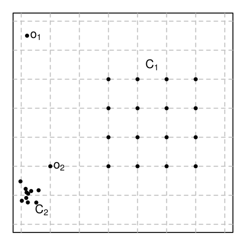
Proximity-based Outlier Detection: Density-based Techniques (cont.)

- LOF: Local Outlier Factor [Breunig et al., 2000]
  - *k-distance*: distance between *p* and its *k*-th nearest neighbour
  - *k-distance neighborhood*: all the points whose distance from *p* is not greater than the *k*-distance.
  - *reachability-distance* of *p* with respect to *o*: the maximum between their *k*-distance and their actual distance.



  - intuition: high values of reachability-distance between two given points indicates that they may not be in the same cluster

- LOF: Local Outlier Factor [Breunig et al., 2000] (cont.)
  - *local reachability-density* of a point is defined to be inversely proportional to the average reachability-distance of its k neighbourhood.
  - *LOF* assigns high values to the points that have a much lower *local reachability-density* in comparison to its k-neighbourhood.
  - Example:



  - o_2 is assigned an higher LOF compared to the the LOF values assigned to the points of C_1 and C_2
  - This captures a local outlier whose local density is relatively low comparing to the local densities of its k-neighbourhood..

Proximity-based Outlier Detection: Density-based Techniques (cont.)

- Multi-granularity Deviation Factor [Papadimitriou et al., 2003]
  - finds not only outlier instances, but groups of outliers, i.e. micro-clusters
- RDF: Relative-Density Factor [Wang et al., 2004]
  - uses a vertical data structure (P-trees) to efficiently index data and prune the points which are deep in clusters, and then detects outliers only within the remaining small subset of the data

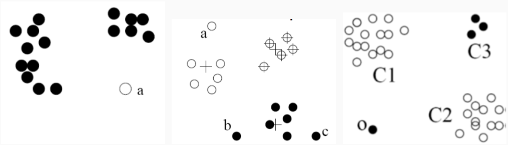
Disadvantages

- True outliers and noisy regions of low density may be hard to distinguish.
- These methods need to combine global and local analysis.
- In high dimensional data, the contrast in the distances is lost.
- Computational complexity of the test phase.

Clustering-based Outlier Detection

Proposal

- Normal instances belong to large and dense clusters, while outlier instances are instances that:
  - do not belong to any of the clusters;
  - are far from its closest cluster;
  - form very small or low density clusters.

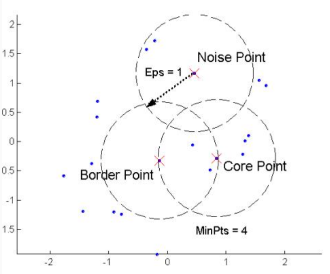


Advantages

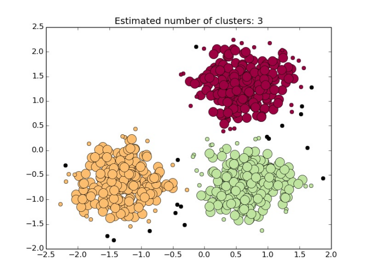
- Easily adaptable to on-line/incremental mode.
- Test phase is fast.

- DBSCAN [Ester et al., 1996]
  - Clustering method based on the notion of “density” of the points
  - The density of a point is estimated by the number of points that are within a certain radius.
  - Based on this idea, points can be classified as:

- *core points*: if the number of points within its radius are above a threshold
- *border points*: if the number of points within its radius are not above a threshold, but they are within a radius of a *core point*
- *noise points*: if do not have enough points within their radius, nor are sufficiently close to any *core point*.

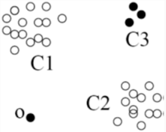


- DBSCAN [Ester et al., 1996] (cont.)
  - *noise points* are removed for the formation of clusters
  - all *core points* that are within a certain distance of each other are allocated to the same cluster
  - each *border point* is allocated to the cluster of the nearest *core points*
  - *noise points* are identified as outliers.



Clustering-based Outlier Detection: Techniques (cont.)

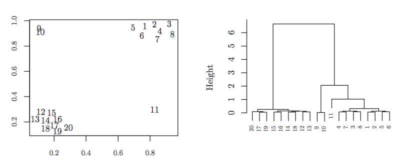
- FindCBLOF [He et al., 2003]
  - Find clusters, and sort them in decreasing order
  - To each point, assign a *cluster-based local outlier factor* (CBLOF)
  - The CBLOF score of a point p is determined by the size of the cluster to which p belongs, and the distance between p and
    - its cluster centroid, if p belongs to a large cluster
    - its closest large cluster centroid, if p belongs to a small cluster.
  - the distance between the point and the cluster, can be the similarity measure used in the clustering algorithm.
  - Example:



- o is outlier since its closest large cluster is C_1, and 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 the size of C_3 is small

Clustering-based Outlier Detection: Techniques (cont.)

- *OR_H* [Torgo, 2007]
  - Obtain an agglomerative hierarchical clustering of the data set
  - Use the information on the "path" of each point through the dendogram as a form to determine its degree of outlyingness
  - Cases that are only merged at later stages are surely very different from others
  - The outlier score of a point is given by the later stage of its agglomerative process and can be estimated by the size difference between the clusters being merged at that stage.
  - The higher the clusters size difference, the higher the outlier score.



Clustering-based Outlier Detection

Disadvantages

- Computationally expensive in the training phase.
- If normal points do not create any clusters, this technique may fail.
- In high dimensional spaces, clustering algorithms may not give any meaningful clusters.
- Some techniques detect outliers as a byproduct, i.e. they are not optimized to find outliers, their main aim is to find clusters.

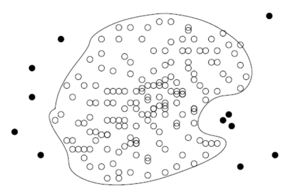
Outlier Detection Approaches

Semi-supervised Learning Techniques

One Class Classification

Proposal

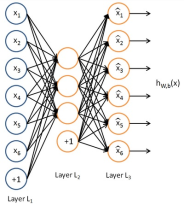
- Build a prediction model to the normal behaviour and classify any deviations from this behaviour as outliers.



Advantages

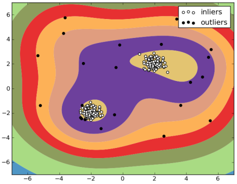
- Models are interpretable.
- Normal behaviour can be accurately learned.
- Can detect new outliers that may not appear close to any outlier points in the training set.

- Auto-associative neural networks [Japkowicz et al., 1995]
  - A feed-forward perceptron-based network is trained with normal data only.
  - The network has the same number of input and output nodes and a decreased number of hidden nodes to induce a bottleneck.
  - This bottleneck reduces the redundancies and focus on the key attributes of data.
  - After training, the output nodes recreate the example given as input nodes.
  - The network will successfully recreate normal data but will generate a high-recreation error for outlier data.



One Class Classification: Techniques (cont.)

- One-class SVM [Tax and Duin, 2004]
  - It obtains a spherical boundary, in the feature space, around the normal data. The volume of this hypersphere is minimized, to minimize the effect of incorporating outliers in the solution.
  - The resulting hypersphere is characterized by a centre **c** and a radius R.
  - The optimization problem consists of minimizing the volume of the hypersphere, so that includes all the training points.
  - Every point lying outside this hypersphere is an outlier.



One Class Classification

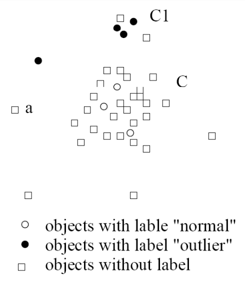
Disadvantages

- Requires previous labeled instances for normal behaviour.
- Possible high false alarm rate - previously unseen normal data may be identified as an outlier.

Hybrid Learning Techniques

For example, combining classification and clustering [Han et al., 2011]

- With some objects labeled as either “normal” or “outlier”
- Using a clustering-based approach, we find a large cluster,C, and a small cluster, C1.
- Because 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.
- Any object that does not fall into the model for C (such as a) is considered an outlier as well.
- Since some objects in cluster C1 carry the label “outlier”, declare all objects in C1 as outliers.



Outlier Detection Approaches

Supervised Learning Techniques

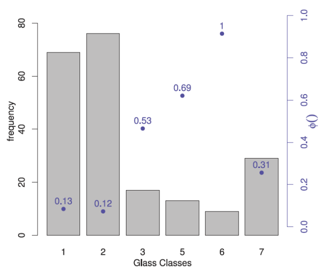
Learning from Imbalanced Domains

- In a supervised learning task the goal is to build a model of an unknown function $Y = f(X_1, X_2, \cdots, X_p)$, based on a training sample $\{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$ with examples of this function.
- Depending on the type of target variable Y, we have:
  - classification task, if Y is nominal
  - regression task, if Y is numeric
- The goal of outlier detection in supervised learning is to learn from a set of cases for which the target variable Y value have poor representativeness on the training data but which are the most relevant ones for the end user.

Learning from Imbalanced Domains (cont.)

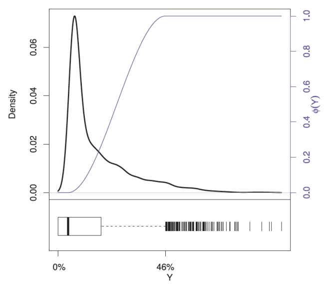
- Classification

  outliers are the cases labeled with infrequent classes in the target variable



- Regression

  outliers are the cases which take values in ranges of the target variable where values are rare



$\phi(Y)$ is a relevance function that maps the values of the target variable Y into a range [0, 1] of importance (1 is the maximal importance)

Learning from Imbalanced Domains (cont.)

- It is of key importance that the obtained models are particularly accurate at the sub-range of the domain of the target variable for which training examples are rare.
- To prevent the models of being biased to the most frequent cases, it is necessary to use:
  - performance metrics biased towards the performance on these rare cases;
  - learning strategies that focus on these rare cases.

Suitable Performance Metrics

Classification

- In a classification setting, this type of problems is usually represented by a 2-class problem where outliers are the minority (positive) class.

2-class Confusion Matrix				
		True		
		Negative	Positive	Total
Predicted	Negative	TN	FP	PNEG
	Positive	FN	TP	PPOS
	Total	NEG	POS	

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

- Standard performance metrics (e.g. *accuracy*) are not suitable for this type of problems.

Classification (cont).

- Example: Diagnose of a rare disease

Model B Confusion Matrix

		Disease: absent	Disease: present
Diagnose	negative	TN = 63	FN = 2
	positive	FP = 27	TP = 8

Model C Confusion Matrix

		Disease: absent	Disease: present
Diagnose	negative	TN = 68	FN = 7
	positive	FP = 22	TP = 3

- The accuracy for both models is 71%.
- Model B correctly diagnosed 80% of the sick individuals
- Model C diagnosed only 30%
- The goal is to achieve a good performance on the outlier cases.

Suitable Performance Metrics (cont.)

Classification: some suitable performance metrics [Branco et al., 2016]

- Precision: ratio between the number of correctly predicted outliers and the total number of cases predicted as outliers. $(TP/(TP + FP))$
- Recall: ratio between the number of correctly predicted outliers and the total number of existing outliers. $(TP/(TP + FN))$
- False Alarm Rate: ratio between the number of normal cases wrongly predicted as outliers and the total number of normal cases. $(FP/(TN + FP))$
- F-measure: trade-off measure between precision and recall.
- ROC Curve and AUC: trade-off between recall and false alarm rate as the discrimination threshold for the two classes is varied.
- PR Curve and AUC-PR: trade-off between recall and precision as the discrimination threshold for the two classes is varied.

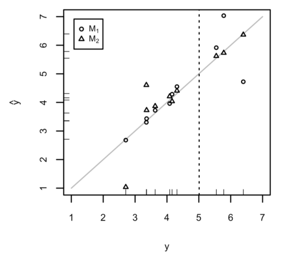
Suitable Performance Metrics (cont.)

Regression

- One of the most commonly used performance metrics in regression is

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Example: Prediction of NO_2 emissions



- Both M_1 and M_2 models achieve an MSE of 0.460
- Still, M_2 is more accurate at higher NO_2 concentration values, the most important to predict accurately.

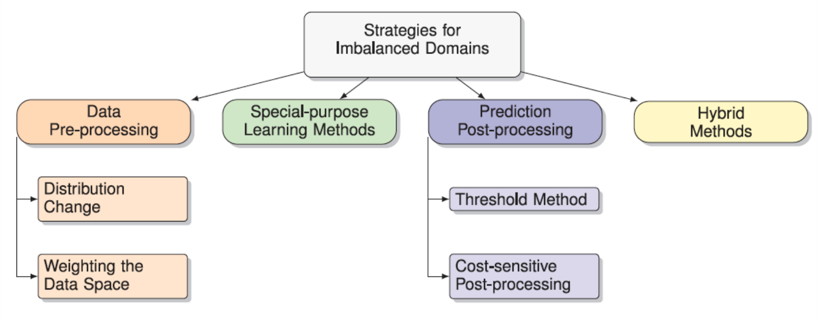
- As in classification, standard performance metrics fail the goal
- The goal is to achieve a good performance on the outlier cases.

Suitable Performance Metrics (cont.)

Regression: some suitable performance metrics [Branco et al., 2016]

- RROC Curve: trade-off between over-estimation and under-estimation errors by varying a shift added/subtracted to the predictions.
- REC Curve: the predictive performance of a regression model across the range of possible errors.
- REC Surfaces: incorporate the cumulative distribution of the target variable in REC Curves to give an insight about the error location across target variable domain.
- MU (mean utility): evaluates the utility of the regression model by taking into account the loss and the relevance of the values involved in each prediction.
- Adaptation of some classification metrics: precision, recall and derived metrics.

Learning Strategies for Imbalanced Domains



Data Pre-Processing Strategies

Proposal

Change the data distribution to make standard algorithm focus on rare and relevant cases.

Advantages

- They allow the application of any learning algorithm
- The obtained model will be biased to the goals of the user
- Models will be interpretable

Data Pre-Processing Strategies (cont.)

Techniques

- Distribution Change
  - change the data distribution with the goal of addressing the issue of poor representativeness of the more relevant cases
- Weighting the data space
  - some algorithms allow different weights to be assigned to different data instances.

Data Pre-Processing Strategies (cont.)

Some Distribution Change Techniques

- Random under-sampling
  - removes examples from the majority class or with common values from the original dataset, reducing the size of the dataset.
  - Problem: useful examples for the learning task may be discarded
- Random over-sampling
  - a random set of copies of minority class or rare values examples is added to the dataset.
  - Problem: possible over-fitting, i.e. poor generalization ability of the model

Some Distribution Change Techniques (cont.)

- SMOTE (Synthetic minority over-sampling technique) [Chawla et al., 2002], SmoteR [Torgo et al., 2013] and other SMOTE variants
  - it over-samples the minority class (or rare values) examples by generating new synthetic data combined with some percentage of random under-sampling of the majority class (common values) examples;
  - the generation of synthetic data reduces the risks of under-sampling and over-sampling;
  - creates new examples by introducing perturbation on the examples or using interpolation of existing examples.

Data Pre-Processing Strategies (cont.)

Disadvantages

- difficulty of relating the modifications in the data distribution and the user preferences
- mapping the given data distribution into an optimal new distribution according to the user goals is not easy

Special-purpose Learning Strategies

Proposal

Change the learning algorithms so they can learn from imbalance data

Advantages

- The user goals are incorporated directly in to the models by setting an appropriate preference criterion.
- Models will be interpretable for the user

Special-purpose Learning Strategies (cont.)

Some Techniques

- RareBoost [Joshi et al., 2001]
  - an ensemble strategy
  - examples of the minority class that are misclassified are assigned higher weights in the next iteration
- PNrule [Joshi et al., 2002]
  - a two-phase rule induction algorithm for classification;
  - P phase covers as many as positive examples as possible (good recall)
  - N phase removes FP, focus on precision.
- ubaRules [Ribeiro, 2011]
  - an ensemble strategy that generates several regression trees
  - selection of some of derived rules into a final ensemble according to a specific preference criterion which maximizes utility.

Special-purpose Learning Strategies (cont.)

Disadvantages

- The user will be restricted to that specific set of modified learning algorithms
- It requires a deep knowledge of the algorithms
- If the preference criterion changes, models have to be relearned and, possibly the algorithm has to be re-adapted
- Is not easy to map the user preferences with a suitable preference criterion

Proposal

Use the original dataset and a standard learning algorithm, only manipulating the predictions of the models according to the user preferences and the imbalance of the data

Advantages

- It is not necessary to be aware of the user preferences at learning time
- The same model can be applied to different deployment scenarios without having to be relearned
- Any standard learning algorithm can be used

Prediction Post-processing Strategies (cont.)

Techniques

- Threshold Method
  - obtain several models by varying the threshold on the score that expresses the degree to which an example is member of a class (e.g. [Weiss, 2004])
- Cost-Sensitive Post-Processing
  - change the model predictions to make it cost-sensitive or to adapt it to a different operating context (e.g. [Hernández-Orallo, 2014])

Prediction Post-processing Strategies (cont.)

Disadvantages

- the models do not reflect the user preferences
- models interpretability is jeopardized as they were obtained by optimzing a function that is not in accordance with the user preference bias

Supervised Learning for Outlier Detection: Wrap-up

Proposal

Build a prediction model for normal and rare classes (values) of the target variable.

Disadvantages

- Has to handle a training set with an imbalanced distribution.
- In classification relies on the availability of accurate labels for the training instances.
- In regression it assumes that the distribution given in the training data is representative and, thus, is not expected to change in the test data.
- Cannot detect unknown or emerging outliers.

Outlier Detection Approaches

Advanced Topics

Proposal

- If a data instance is an outlier in a specific context (but not otherwise), then it is considered as a contextual outlier.

- Each data instance is defined using two sets of attributes:
  - Contextual attributes used to determine the context (or neighbourhood) for that instance.
    - Sequential Context: position, time.
    - Spatial Context: latitude, longitude.
    - Graph Context: weights, edges.
  - Behavioural attributes which define the non-contextual characteristics of an instance.

- The outlier behaviour is determined using the values for the behavioural attributes within a specific context.

Contextual Outlier Detection (cont.)

Example:

- Detect outlier customers in the context of customer groups
  - Contextual attributes: age group, postal code
  - Behavioural attributes: the number of transactions per year, annual total transaction amount

Advantages

- Allow a natural definition of outlier in many real-life applications.
- Detects outliers that are hard to detect when analyzed in the global perspective.

Contextual Outlier Detection (cont.)

Techniques

- Reduction to point outlier detection
  - Segment data using contextual attributes.
  - Apply a traditional point outlier within each context using behavioural attributes.
  - Model “normal” behaviour with respect to contexts: an object is an outlier if its behavioural attributes significantly deviate from the values predicted by the model.
- Utilizing structure in data
  - Build models from the data using contextual attributes to predict the expected behaviour with respect to a given context.
  - Avoids explicit identification of specific contexts

Disadvantages

- Identifying a set of good contextual attributes.
- It assumes that all normal instances within a context will be similar (in terms of behavioural attributes), while the outliers will be different.

Collective Outlier Detection

Proposal

- If a collection of related data instances is anomalous with respect to the entire data set, then it is considered a collective outlier.
- The individual data instances in a collective outlier may not be outliers by themselves, but their occurrence together as a collection is anomalous.

Advantages

- Allow a natural definition of outlier in many real-life applications in which data instances are related.

Collective Outlier Detection (cont.)

Techniques

- A collective outlier can also be a contextual outlier if analyzed with respect to a context.
- A collective outlier detection problem can be transformed to a contextual outlier detection problem by incorporating the context information.

Collective Outlier Detection (cont.)

Disadvantages

- Contrary to contextual outliers, the structures are often not explicitly defined, and have to be discovered as part of the outlier detection process.
- Need to extract features by examining the structure of the dataset, i.e. the relationship among data instances for:
  - sequence data to detect anomalous sequences;
  - spatial data to detect anomalous sub-regions;
  - graph data to detect anomalous sub-graphs.
- The exploration of structures in data typically uses heuristics, and thus may be application dependent.
- The computational cost is often high due to the sophisticated mining process.

Outlier Detection in High Dimensional Data

Challenges

- Interpretation of outliers
  - Detecting outliers without saying why they are outliers is not very useful in high-D due to the many features (or dimensions) involved
  - Identify the subspaces that manifest the outliers
- Data sparsity
  - Data in high-D spaces is often sparse
  - The distance between objects becomes heavily dominated by noise as the dimensionality increases
- Data subspaces
  - Capturing the local behavior of data
- Scalable with respect to dimensionality
  - # of subspaces increases exponentially

Outlier Detection in High Dimensional Data (cont.)

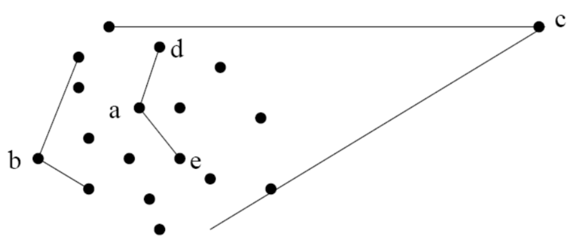
Techniques

- Find distance-based outliers, but use the ranks of distance instead of the absolute distance in outlier detection.
- Dimensionality reduction: 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.
- Project data onto various subspaces to find an area whose density is much lower than average.

Techniques (cont.)

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

  E.g. Angle-based outliers.



Summary

Summary

- Outliers are not necessarily random noise.
- They can represent critical information that can trigger preventive or corrective actions.
- The interpretability of an outlier detection method is extremely important.
- The nature of the outlier detection problem is dependent on the application domain.
- Different approaches to this problem are necessary.
- Contextual and collective outliers are having increasing applicability in several real-world domains.
- Online Outlier Detection and Distributed Outlier Detection are emerging topics.
- There is much space for the development of new techniques in this area.

References

References

Aggarwal, C. (2013).
***Outlier Analysis*.**
Springer New York.

Aggarwal, C. C. (2015).
***Data Mining, The Texbook*.**
Springer.

Branco, P., Torgo, L., and Ribeiro, R. P. (2016).
A survey of predictive modeling on imbalanced domains.
ACM Comput. Surv., 49(2):31:1–31:50.

Breunig, M. M., Kriegel, H. P., Ng, R., and Sander, J. (2000).
Lof: Identifying density-based local outliers.
In *Proceedings of ACM SIGMOD 2000 International Conference on Management of Data*.
ACM Press.

Chandola, V., Banerjee, A., and Kumar, V. (2009).
Anomaly detection: A survey.
ACM Computing Surveys (CSUR), 41(3):15.

Chawla, N. V., Bowyer, K. W., Hall, O. L., , and Kegelmeyer, W. P. (2002).
Smote: Synthetic minority over-sampling technique.
Journal of Artificial Intelligence Research, 16:321–357.
AAAI Press.

Ester, M., peter Kriegel, H., S, J., and Xu, X. (1996).
A density-based algorithm for discovering clusters in large spatial databases with noise.
pages 226–231. AAAI Press.

Fan, W., Stolfo, S., Zhang, J., and Chan, P. K. (1999).
Adacost: Misclassification cost-sensitive boosting.
In *ICML'99: Proceedings of 16th International Conference on Machine Learning*, pages 97–105. Morgan Kaufmann Publishers Inc.

Han, J., Kamber, M., and Pei, J. (2011).
***Data Mining: Concepts and Techniques*.**
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Hawkins, D. M. (1980).
***Identification of Outliers*.**
Chapman and Hall.

References (cont.)

He, Z., Xu, X., and Deng, S. (2003).
Discovering cluster based local outliers.
Pattern Recognition Letters, 2003:9–10.

Hempstalk, K., Frank, E., and Witten, I. H. (2008).
One-class classification by combining density and class probability estimation.
In *ECML/PKDD (1)*, pages 505–519.

Hernández-Orallo, J. (2014).
Probabilistic reframing for cost-sensitive regression.
ACM Trans. Knowl. Discov. Data, 8(4):17:1–17:55.

Hodge, V. J. and Austin, J. (2004).
A survey of outlier detection methodologies.
Artificial Intelligence Review, 22:2004.

Japkowicz, N., Myers, C., and Gluck, M. A. (1995).
A novelty detection approach to classification.
In *IJCAI*, pages 518–523. Morgan Kaufmann.

References (cont.)

Joshi, M. V., Agarwal, R. C., and Kumar, V. (2002).
Predicting rare classes: Comparing two-phase rule induction to cost-sensitive boosting.
In *PKDD'02: Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 2431 of *LNCS*, pages 237–249. Springer.

Joshi, M. V., Kumar, V., and Agarwal, R. C. (2001).
Evaluating boosting algorithms to classify rare classes: Comparison and improvements.
In *Proceedings of the 2001 IEEE International Conference on Data Mining, 29 November - 2 December 2001, San Jose, California, USA*, pages 257–264.

Knorr, E. M. and Ng, R. T. (1998).
Algorithms for mining distance-based outliers in large datasets.
In *VLDB'98: Proceedings of 24th International Conference on Very Large Data Bases*, pages 392–403. Morgan Kaufmann, San Francisco, CA.

Kubat, M. and Matwin, S. (1997).
Addressing the curse of imbalanced training sets: one-sided selection.
In *Proc. 14th International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann.

References (cont.)

Lazarevic, A. (2008).
Anomaly detection: A tutorial.
Tutorial Session on 2008 Siam Conference on Data Mining (SDM08).

Maloof, M. A. (2003).
Learning when data sets are imbalanced and when costs are unequal and unknown.
In *ICML-2003 workshop on learning from imbalanced data sets II*, volume 2, pages 2–1.

Papadimitriou, S., Kitagawa, H., Faloutsos, C., and Gibbons, P. B. (2003).
Loci: Fast outlier detection using the local correlation integral.
In *ICDE'03: Proceedings of 19th International Conference on Data Engineering*, pages 315–326. IEEE Computer Society.

Ribeiro, R. P. (2011).
***Utility-based Regression*.**
PhD thesis, Dep. Computer Science, Faculty of Sciences - University of Porto.

Tax, D. (2001).
***One-class classification: Concept learning in the absence of counter-examples*.**
PhD thesis, Technische Universiteit Delft.

Tax, D. M. J. and Duin, R. P. W. (2004).
Support vector data description.
Machine Learning, 54(1):45–66.

References (cont.)

Torgo, L. (2007).
Resource-bounded fraud detection.
In *Progress in Artificial Intelligence, 13th Portuguese Conference on Aritficial Intelligence, EPIA 2007, Workshops*, pages 449–460.

Torgo, L. (2016).
Outlier detection methods.
Slides.

Torgo, L. (2017).
***Data Mining with R: Learning with Case Studies*.**
Chapman and Hall/CRC, 2nd edition.

Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P. (2013).
Smote for regression.
In *Progress in Artificial Intelligence*, pages 378–389. Springer.

Wang, B., Ren, D., and Perrizo, W. (2004).
Rdf: A density-based outlier detection method using vertical data representation.
In *13th International Conference on Data Mining (2004)*, pages 503–506. IEEE.

Weiss, G. M. (2004).
Mining with rarity: a unifying framework.
SIGKDD Explorations Newsletter, 6(1):7–19.

References (cont.)

Yu, D., Sheikholeslami, G., and Zhang, A. (2002).
Findout: Finding outliers in very large datasets.
Knowledge and Information Systems, 4(4):387–412.

Zhang, Y., Meratnia, N., and Havinga, P. (2007).
A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets.