FACULTY
OF INFORMATICS
Masaryk University

Ensembles For Anomaly Detection

A Brief Introduction

Tomas Kruty

Outline for Section 1

Motivation

- Data cleansing
- Fraud detection
- Interesting event detection
- Medical diagnosis
- Law enforcement

Local Outlier Factor

- Based on local density of observation's neighborhood
- Core distance of point p
  - distance between p and it's $k'th$ nearest neighbor
- Reachability distance between observations $p1$ and $p2$
  - maximum of core distance of $p1$ and the distance between $p1$ and $p2$
- Local reachability distance
  - inversely proportional to the average reachability distance of its k neighbors
- LOF of an observation is calculated as a function of it's local reachability distance

Clustering-Based Outlier Rankings

- uses a hierarchical agglomerative clustering algorithm
- uses the information about clustering process to determine outliers
- outliers will be more resistant to being merged than other observations
- the size difference between the group to which the outlier belongs and to which is being merged should be very large

Class Outliers: Distance-Based

- Class Outlier Mining
  - given a set of observations with class labels, find those that arouse suspicions, taking into account the class labels
- The Probability of the class label
  - the probability of the class label of the instance T with respect to the class labels of its K Nearest Neighbors
- Deviation
  - how much the instance *T* deviates from subset of observation with same class as *T*
- K-Distance
  - distance between the instance *T* and its *K* nearest neighbors
- $COF = K * PCL + (\alpha/Deviation) + (\beta * \text{K-Distance})$

Outline for Section 2

General Idea

- Method of combining multiple different algorithms (or different instances of an algorithm)
- Ensembles should provide more robust results
- Ensemble is responsible for combining the outputs of algorithms used in the ensemble

Categorization

by component independence

- Sequential ensembles
  - set of algorithms are applied sequentially
  - future applications of algorithm are affected by previous application
  - result is either weighted combination of or last application of outlier analysis application
- Independent ensembles
  - different algorithms are applied to data
  - results obtained from different applications are independent
  - outputs from different algorithms are combined together for more robust outliers

Categorization

by component

- Voting-based
  - different algorithms vote on the output of the ensemble
  - simple approach is to assign each algorithm one vote but prioritization is possible
- Bagging-based
  - bootstrap aggregation or bagging for short
  - uses only one algorithm but multiple times each on different subset of the data - so called bags
  - bags can be created by taking a subset of observations or a subset of features
- Model-based
  - uses one dataset but multiple algorithms
  - the challenge of combining various outputs
  - normalization or outputs must take place

LOF Ensembles

- Building the ensemble:
  - different values of K
- Combining the outputs:
  - mean of outlier scores
  - maximum of outlier scores
  - voting on observations
  - weighted average of outlier scores

OR_h Ensembles

- Building the ensemble
  - different algorithms for clustering
  - different method for obtaining outlier score
  - bagging
- Combining the outputs
  - mean of outlyingness factor
  - maximum of outlyingness factor

CODB Ensembles

- Building the ensemble
  - different values of K
- Combining the outputs
  - mean of outlier scores
  - minimum of outlier scores

RF-OEX

- Ensemble method based on random forests
- Proximity to the members of the same class
  - inverse value is used, the higher proximity to observation's class the less of an outlier the given observation is
  - proximity to class C is computed as an aggregation of proximities to all observations from C
- Misclassification measure
  - similarity with members of different classes should increase observation's outlyingness
  - number of observations with different class in analyzed observation's close proximity
- Ambiguity measure
  - increase of importance of outliers that are far from all observations
- $OF(p) = OF_1(p)_{same-class} + OF_2(p)_{misclassification} + OF_3(p)_{ambiguity}$

Bibliography

[1] Charu C. Aggarwal. *Outlier Analysis*. Springer International Publishing, 2016.

[2] Luis Torgo. *Data Mining with R: Learning with Case Studies, Second Edition*. CRC Press, 2016.

[3] N. Hewahi, and M. Saad. *Class outliers mining: Distance-based approach*. International Journal of Computer and Information Engineering, 2007.

[4] Charu C. Aggarwal. *Outlier Ensembles: An Introduction*. Springer International Publishing, 2017.

[5] Leona Nezvalova, Lubos Popelinsky, Luis Torgo, and Karel Vaculik. *Class-based outlier detection: staying zombies or awaiting for resurrection?*. KD Lab, FI MU BRNO, and F.Sci. U.Porto.