Charu C. Aggarwal
IBM T J Watson Research Center
Yorktown, NY 10598

Outlier Ensembles

Keynote, Outlier Detection and Description Workshop, 2013

Based on the ACM SIGKDD Explorations Position Paper: "Outlier Ensembles"

Introduction

- The objective function or model for a data mining problem is often constructed using a subjective and heuristic process based on an analyst's understanding.
  - Should an outlier be distance-based, linear-model based or probabilistic?
  - Such assumptions can often be imperfect, and a specific algorithm being used may model the underlying generative process in a limited way.
- Because of this imperfection, a model may work better on some parts of the data than other.
- Similarly, a given model may be better than another in a data-specific way, which is unknown a-priori.

Ensemble Analysis

- Ensemble analysis is a method which is commonly used in the literature in order to reduce the dependence of the model on the specific data set or data locality.
  - Greatly increases the robustness of the data mining process.

- The ensemble technique is used very commonly in problems such as clustering and classification.

- **Broad Idea:** Combine the results from different models in order to create a more robust model.
  - Tremendous variation in how the different models are selected and combined.

Example: Classification and Clustering

- *Heterogeneous Model Averaging:* Construct different classes of models (eg. decision trees, rules, Bayes) or many instantiations of the same class, and vote on the class label of a test instance.

- *Bagging:* Sample repeatedly from training data, and vote on the class label of a test instance.

- *Boosting;* Sequentially select more "difficult" subsets of the training data, and use a weighted combination of votes on the test instance.

- *Multiview and Alternative Clustering:* Construct clusterings which are orthogonal to one another by using techniques such as spectral methods, and combine results from different instantiations.

Relative Status of Methods for Outlier Analysis

- The problem of ensemble analysis has been widely studied in the context of problems such as clustering and classification.
  - Each of these areas of meta-algorithm analysis is considered an active and vibrant subfield in its own right.
  - Eg. The seminal paper on boosting in classification has several thousand citations.
- Remotely not true for outlier analysis, in which the work on ensemble analysis is rather patchy, sporadic, and not so well formalized.
- In many cases, useful meta-algorithms are buried deep into the algorithm, and not formally recognized as ensembles.

Challenges in Outlier Analysis

- Ensemble analysis is generally more difficult in the context of outlier detection.
  - *Unsupervised Nature:* Crisp evaluation criteria are useful in ensemble techniques such as boosting, where sequential analysis is used.
    - Classification has a richer ensemble literature as compared to clustering
  - *Small Sample Space Problem:* A given data set may contain only a small number of outliers.
    - Even harder to quantify approach robustly.
    - Problem for making robust decisions about future steps of the algorithm, without overfitting.
    - Unique problem in outlier analysis.

Current Status

- Ensemble analysis has currently started receiving attention in the outlier analysis literature.

- A particular case where ensemble analysis is commonly used is that of high dimensional data.

  – Earliest *formalization* of outlier ensemble analysis was a feature bagging approach used in high dimensional outlier detection (Lazarevic et al).

  – Most current applications of ensemble analysis are designed for the high dimensional scenario, though *potential applicability* is much broader.

Application to High Dimensional Outlier Analysis

- High dimensional scenario is an important one for ensemble analysis.
  - The outlier behavior of a data point in high dimensional space is often described by a subset of dimensions.
  - The dimension subsets are rather hard to discover in real settings.
  - Most methods for localizing the subsets of dimensions can be considered *weak guesses* to the true subsets of dimensions which are relevant for outlier analysis.
- The ensemble approach improves the robustness and uncertainty of the results obtained from the subspace discovery process.

Historical Perspective

- The feature bagging work discussed in Lazarevic et al may be considered a first *formal* description of outlier ensemble analysis in a real setting.

- Numerous methods were proposed earlier to this work which could be considered ensembles, but were never formally recognized as ensembles in the literature.

- Even automated parameter tuning methods in some classical outlier detection methods (eg. LOF) are typically structured as ensemble methods.

  – While these papers have implicitly used the insight of ensemble analysis, the papers did not focus on claiming the idea as a general meta-algorithm!

Example: LOF

- LOF quantifies the local density of a data point, with the use of a neighborhood of size k.

- How to pick the value of k?

- Apply the algorithm over different values of k and pick the value of k which provides the strongest outlier score $\Rightarrow$ Ensemble Analysis!

  – An advantage of LOF is that scores are normalized, which means that values can be compared over different values of k.

  – Not true across all algorithms; eg. trying to compare k-nearest neighbor distance scores, across different functions or dimensionalities $\Rightarrow$ Normalization is important.

Example: LOCI

- LOCI computes densities in the neighborhood as well, except that it uses a *sampling neighborhood radius* and a *counting neighborhood radius*, which are related to one another by a constant factor.

- How to compute appropriate neighborhood size? $\Rightarrow$ Multi-granularity approach over different radius sizes, and pick strongest score.

- LOCI plot pictorially illustrates the outlier behavior over different components of the ensemble.
  - Provides excellent visual interpretability $\Rightarrow$ Relevant to outlier description.

Feature Bagging

- Paper provides first formal description as a general purpose meta-algorithm.

- Randomly sample subspaces of dimensionality between $d/2$ and d, and compute LOF outlier score.

- Compute highest score across all subspaces
  - Another combination variant uses averaging across samples

Basic Ensemble Algorithm

- Derive different outlier scores for a data point using different methods, data selection schemes etc.

  - The different outlier scores may be derived using schemes which are either independent of one another or dependent on one another.

- Combine scores from different algorithms to obtain (a more robust) outlier score.

Key Challenges

- How to design the ensemble?
  - Choice of models and dependency of models
- What if scores cannot be meaningfully compared with one another?
- One outlier score may use a maximization objective, and another might use a minimization objective
  - Normalization is important!
- How to combine? $\Rightarrow$ Average, Maximum?

Different Types of Ensembles

- Independent Ensembles vs Sequential Ensembles
  - Are the components designed independent of one another or dependent on each other (eg. successive refinement)?

- Model-centered vs. Data-centered
  - Do the different components depend on different outlier detection algorithms or the same algorithms on different derivatives from the data?

Independent vs Sequential Ensembles

- In independent ensembles, independent models are constructed from the data, and combination is used.
  - Most common approach for ensemble analysis.
  - Simple approach in terms of implementation.

- In sequential ensembles, models are successively refined.
  - Advantage of using insights from the previous execution to further refine the model.
  - Unsupervised nature (lack of ground truth) makes refinement a challenge $\Rightarrow$ Rough outlier score-based refinement rather than ground-truth based refinement (as in boosting).

Implementation Differences

- Independent Ensemble: Repeated *Independent* Execution and Combination of Scores: (iteration j)

  Pick an algorithm $\mathcal{A}_j$;

  Create a new data set $f_j(\mathcal{D})$ from $\mathcal{D}$;

  Apply $\mathcal{A}_j$ to $f_j(\mathcal{D})$;

- Sequential Ensemble: Repeated *Sequential* Execution and Combination of Scores: (iteration j)

  Pick an algorithm $\mathcal{A}_j$ based on results from past executions;

  Create a new data set $f_j(\mathcal{D})$ from $\mathcal{D}$ from past execution results;

  Apply $\mathcal{A}_j$ to $f_j(\mathcal{D})$;

Examples

- **Feature Bagging:** Uses independent executions of LOF algorithm on different subspaces to combine scores $\Rightarrow$ Independent

- **OUTRES:** Recursive exploration of subspaces (dependent) and combination of outlier score $\Rightarrow$ Sequential

- **Barbara et al SAC'03, Bootstrapping an intrusion detection system:** Successively remove data points with high outlier score. $\Rightarrow$ Sequential
  - In sequential ensembles, only score based refinement can be used, rather than ground-truth based, which is rather rough.
  - Sequential Ensembles are less common.

Model-Centered vs. Data-Centered

- In model-centered ensembles, different models (possibly same algorithm with different parameter settings) may be applied.

- In data-centered ensembles, same algorithm may be applied to different *derivations* (eg. subsets, subspaces) of the data.

- Possible to create heterogeneous models containing both.

- Distinction between the two is a bit artificial:

  – A data-centered ensemble can be considered a model-centered ensemble by incorporating a data-derivation pre-processing phase.

  – Distinction useful for conceptual design process.

Examples

- **Feature Bagging:** Data-centered ensemble, since it samples subspaces of the data.

- **OUTRES:** Data centered ensemble for same reason as above.

- **LOF-Tuning:** Model-centered ensemble, because it uses the same data, with different parameter settings from the same algorithm.

Heterogeneity Issues

- Possible to combine data- and model-centered ensembles

- Since scores are combined together, the scores from different algorithms may not be meaningfully comparable.

- How to combine an LOF score with a k-nearest neighbor score?

- What if one outlier model works with a score maximization formulation, and another works with a minimization formulation?

- Relevant to several research issues in ensemble analysis.

Research Issues in Score Combination

- Given a set of scores, how do we combine them together? What combination function should be used?

- Given a set of scores, how do we normalize the scores in order to make them meaningfully comparable?

Normalization Issues

- Crucial to understand the statistical significance of a score.

- Ideally, we would like to measure a score as an intuitive probability value.

- Model scores as a 1-dimensional distribution, and convert to probabilities, by using a simple measure such as CDF of distribution!

- Ordering of scores can be addressed during modeling, and final probabilities can always be expressed in maximization form, irrespective of algorithm.

- *J. Gao and P.-N. Tan. Converting output scores from outlier detection algorithms into probability estimates. ICDM, 2006.*

Combination Issues

- Assume that higher score is better (after normalization).

- Commonly used combination functions:

  - Maximum of constituent scores $\Rightarrow$ If *best* description/causality suggests a strong outlier, then consider it an outlier.

  - Average/Sum of constituent scores $\Rightarrow$ If *many* descriptions/causalities suggest a strong outlier, then consider it an outlier.

  - Product of scores: sum of damped (logarithm of) scores $\Rightarrow$ OUTRES

- Maximum and average are most common.

Maximum or Average for Combination?

- *Average* risks dilution from bad models.

- The use of *maximum* can overestimate (absolute) outlier scores by chance over many ensemble components (Bonferroni Principle).

  - Criticism of *maximum* not valid, because outlier scores are not absolute, but relative.

  - The same Bonferroni correction applies to all the data points, and so *ranking* is robust.

  - The dilution from averages can sometimes be drastic eg. LOF will found bad outliers over many larger values of $k \Rightarrow$ The rare nature of outliers is such that many difficult outliers may not be found by a majority of ensemble components $\Rightarrow$ Irrelevant components dominate.

Other Combination Functions

- Not all constituent components may be treated evenly in analysis.

- Consider a sequential ensemble in which model is successively refined using information from previous iteration.

- Score from *last execution* may be reported.

Characteristics of Some Common Algorithms

Method	Model-Centered or Data-Centered	Sequential or Independent	Combination Function	Normalization
LOF Tuning	Model	Independent	Max	Not Needed
LOCI Tuning	Model	Independent	Max	Not Needed
Feature Bagging	Data	Independent	Max/Avg	No
HICS	Data	Independent	Selective Avg	No
Calib. Bagging	Both	Independent	Max/Avg	Yes
OutRank	Data	Independent	Harmonic Mean	No
Multiple Proclus	Data	Independent	Harmonic Mean	No
Converting scores to probabilities	Both	Independent	Max/Avg	Yes
Intrusion Bootstrap	Data	Sequential	Last Component	Not Needed
OUTRES	Data	Sequential	Product	No
Nguyen et al	Both	Independent	Weighted Avg.	No
Isolation Forest	Model	Independent	Expon. Avg.	Yes

Ideas from Clustering and Classification

- **Boosting from Classification:** Harder to generalize because of lack of ground truth.
  - Broader principles can be used in the context of sequential ensembles
- **Bagging:** Already adapted in the context of subspace sampling (feature bagging).
- **Random Forests:** Adapted as Isolation Forests
- **Bucket of Models:** Adapted regularly in a variety of methods.

Discussion of State-of-the Art

- Most of the current ensemble-based methods are relatively simple techniques

- Numerous ideas can be adapted from the current literature on classification and clustering

  - **Caveat:** No ground truth is available with supervision, and score-based adaptations may need to be used

- Tremendous scope exists for advancement in the area.

Conclusions

- Ensemble analysis is a recently emerging area in outlier analysis.

- Has been studied extensively in the literature, without formal recognition.

- Extensively studied in the context of high dimensional analysis, but potential applicability is much broader.

- Existing literature in classification and clustering provides guidance about development of algorithms in the area.

- Fruitful area for further research, but more challenging than the clustering and classification scenarios.