Classification Based Outlier Detection Techniques

Dr. Shuchita Upadhyaya, Karanjit Singh
Dept. of Computer Science and Applications, Kurukshetra University
Haryana, India
HQ Base Workshop Group EME, Meerut Cantt
UP, India

Abstract — Outlier detection is an important research area forming part of many application domains. Specific application domains call for specific detection techniques, while the more generic ones can be applied in a large number of scenarios with good results. This survey tries to provide a structured and comprehensive overview of the research on Classification Based Outlier Detection listing out various techniques as applicable to our area of research. We have focused on the underlying approach adopted by each technique. We have identified key assumptions, which are used by the techniques to differentiate between normal and Outlier behavior. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain. We provide a basic outlier detection technique, and then show how the different existing techniques in that category are variants of this basic technique. This template provides an easier and succinct understanding of the Classification based techniques. Further we identify the advantages and disadvantages of various classification based techniques. We also provide a discussion on the computational complexity of the techniques since it is an important issue in our application domain. We hope that this survey will provide a better understanding of the different directions in which research has been done on this topic, and how techniques developed in this area can be applied in other domains for which they were not intended to begin with.

Keywords — Outliers, Classification, Outlier Detection, Classification based Outlier Detection, One-Class, Multi-Class, Algorithms, Data Mining.

I. INTRODUCTION

A. General Description and Underlying Assumptions

Classification [1,2] is used to learn a model (classifier) from a set of labeled data instances (training) and then, classify a test instance into one of the classes using the learnt model (testing). Classification based outlier detection techniques operate under the general assumption that a classifier can be learnt from a given feature space that can distinguish between normal and outlier classes.

B. General Methodology of Operation

Classification based outlier detection techniques operate in two phases.

1) Training - The training phase learns a classifier using the available labeled training data.

2) Testing – This phase classifies a test instance as normal or an outlier using the classifier.

II. CATEGORIZATION OF CLASSIFICATION BASED OUTLIER DETECTION TECHNIQUES

These techniques can be grouped into two categories:

A. Multi-Class

These techniques assume that the training data contains labeled instances which belong to multiple normal classes [3,4]. One has to learn a classifier to distinguish between each normal class against the rest of the classes. Refer Figure 1 for illustration. If a test instance is not classified as normal by any of the classifiers, then it is considered as an outlier. Some multi-category techniques give the prediction made by the classifier a confidence score. If none of the classifiers are confident in classifying the test instance as normal or in other words do not score well, the instance is declared to be an outlier.

![Figure 1: Multi-class Classification Based Outlier Detection](http://www.internationaljournalssrg.org)

B. One-Class

One-class classification based outlier detection techniques assume that all training instances bear a single class label. After learning a discriminative boundary around the normal instances using suitable one-class classification algorithm if any test instance does not fall within the learnt boundary it is...
considered an outlier. Some well known algorithms are as listed below:-

1) One-class SVMs [5].
2) One-class Kernel Fisher Discriminants [6,7], as shown in Figure 2.

III. NEURAL NETWORKS BASED TECHNIQUES

A. General Description

Neural networks can be applied to outlier detection in multi-class as well as one-class scenarios. The basic technique of operation of multi-class outlier detection using neural networks operates is again two step.

1) Training : First, in the training phase the neural network is trained on normal training data in order to learn the different normal classes.

2) Testing : Second, wherein each test instance is provided as an input to the neural network. If the test input is accepted by the network, it is considered normal and if the network rejects a test input, it is considered as an outlier [3,8].

B. Variants of the Technique

Several variants of the basic neural network technique have been proposed that use different types of neural networks, as summarized in Table I. These are elaborated subsequently.

C. Replicator Neural Networks

These can be used for one-class outlier detection [Hawkins et al. 2002; Williams et al. 2002]. Using their inputs as outputs internally, they are self organizing, forming a compressed representation for the input data. Usually a multi-layer feed forward neural network is constructed that has the same number of input and output neurons (corresponding to the features in the data). In the training phase data is compressed into three hidden layers. The testing phase involves reconstructing each data instance \( x_i \) using the learnt network to obtain the reconstructed output \( o_i \). The reconstruction error for the test instance \( x_i \) is then computed as

\[
\delta_i = \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - o_{ij})^2
\]

where \( n \) is the number of features over which the data is defined. The reconstruction error is directly used as an outlier score for the test instance.

IV. BAYESIAN NETWORKS BASED TECHNIQUES

A. General Description for Univariate Data Set

Bayesian networks can be used in the multi-class setting for outlier detection. In its basic form for a univariate categorical data set, a naïve Bayesian network estimates the posterior probability of observing a class label (from a set of normal class labels and the outlier class label), given a test data instance. The predicted class for the given test instance is the class label with the largest posterior. The likelihood of observing the test instance, given a class and the prior on the class probabilities, are estimated from the training data set. Laplace Smoothing is used for smoothing the zero probabilities, especially for the outlier class.

B. Extension of the Technique to Multivariate Data

By aggregating the per-attribute posterior probabilities for each test instance and then using this aggregated value to assign a class label to the test instance the same basic technique can be generalized for multivariate categorical data set. The basic technique described above assumes independence between the different attributes. Examples are as tabulated below:-

<table>
<thead>
<tr>
<th>Neural Network Used</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi Layered Perceptrons</td>
<td>[9,10,11,12]</td>
</tr>
<tr>
<td>Radial Basis Function Based</td>
<td>[13,14,15]</td>
</tr>
<tr>
<td>Neural Trees</td>
<td>[16]</td>
</tr>
<tr>
<td>Oscillatory Networks</td>
<td>[17,18]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Application</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Intrusion Detection</td>
<td>[29,30]</td>
</tr>
<tr>
<td>Novelty Detection in Video Surveillance</td>
<td>[31]</td>
</tr>
<tr>
<td>Outlier Detection in Text Data</td>
<td>[32]</td>
</tr>
<tr>
<td>Disease Outbreak Detection</td>
<td>[33]</td>
</tr>
</tbody>
</table>
More complex Bayesian networks such as those proposed in [34, 35, 36] capture the conditional dependencies between the different attributes.

V. SUPPORT VECTOR MACHINES BASED

A. General Description

Support Vector Machines (SVMs) [37] have been applied to outlier detection in the one-class setting. Such techniques use one class learning [38] to learn a region that contains the training data instances (a boundary). For learning complex regions Kernels, such as radial basis function (RBF) kernel, can be used. For each test instance, the basic technique determines if the test instance falls within the learnt region. If a test instance falls within the learnt region, it is declared as normal, else it is declared as an outlier.

B. Variants

<table>
<thead>
<tr>
<th>Application</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novelty Detection in Power Generation Plants</td>
<td>[40]</td>
</tr>
<tr>
<td>System Call Intrusion Detection</td>
<td>[41]</td>
</tr>
<tr>
<td>Outliers in Temporal Sequences</td>
<td>[42]</td>
</tr>
</tbody>
</table>

A variant of the basic technique [43] finds the smallest hyper-sphere in the kernel space, which contains all training instances, and then determines which side of that hyper-sphere does a test instance lie. If a test instance lies outside the hypersphere, it is declared to be an outlier.

Robust Support Vector Machines (RSVM) which are robust to the presence of outliers in the training data have been proposed by [44]. RSVM have also been applied to system call intrusion detection [45].

VI. RULE BASED ASSOCIATION

C. General Description

Rule based outlier detection techniques learn rules that capture the normal behaviour of a system. A test instance not covered by any such rule is considered as an outlier. Such techniques can be applied in one-class as well as multi-class setting.

D. Basic Technique

A basic multi-class rule based technique consists of two steps.

1) Learning Rules from Training Data - Each rule has an associated confidence value. This is proportional to the ratio between the number of training instances correctly classified by the rule and total number of training instances covered by the rule. Algorithms, such as RIPPER, Decision Trees are commonly used for learning.

2) For Each Test Instance Find the Rule that Best Captures the Test Instance - The inverse of the confidence associated with the best rule is the outlier score of the test instance.

TABLE IV

ASSOCIATION RULE BASED OUTLIER DETECTION TECHNIQUES USING MULTIVARIATE DATA

<table>
<thead>
<tr>
<th>Application</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Intrusion Detection</td>
<td>[46,47]</td>
</tr>
<tr>
<td>System Call Intrusion Detection</td>
<td>[48, 49]</td>
</tr>
<tr>
<td>Credit Card Fraud Detection</td>
<td>[51]</td>
</tr>
<tr>
<td>Fraud Detection in Spacecraft House Keeping Data</td>
<td>[52]</td>
</tr>
</tbody>
</table>

Association rules are generated from a categorical data set. To ensure that the rules correspond to strong patterns, a support threshold is used to prune out rules with low support [53]. Association rule mining [54] has been used for one-class outlier detection by generating rules from the data in an unsupervised fashion. In the intermediate step of association rule mining algorithms, frequent item sets are generated. An outlier detection algorithm for categorical data sets has been proposed in which the outlier score of a test instance is equal to the number of frequent item sets it occurs in [54].

VII. COMPUTATIONAL COMPLEXITY

The computational complexity of classification based techniques depends on the classification algorithm being used.

A. Training Phase

The complexity of training classifiers has been covered in [56]. Generally, training decision trees tends to be faster while techniques that involve quadratic optimization, such as SVMs, are more expensive, though some linear time SVMs [57] have been proposed that have linear training time.

B. Testing Phase

The testing phase of classification techniques is usually very fast since the testing phase uses a learnt model for classification.

VIII. ADVANTAGES AND DISADVANTAGES OF CLASSIFICATION BASED TECHNIQUES

A. Advantages

The advantages of classification based techniques are as follows:

1) Use of Powerful Algorithms - Classification based techniques, especially multi-class, can make use of powerful algorithms that can distinguish between instances belonging to different classes.
2) Fast testing Phase - The testing phase is fast since each test instance gets compared against the pre-computed model.

B. Disadvantages

The disadvantages of classification based techniques are as follows:

1) Non-Availability of Accurate Labels for Various Normal classes - Multi-class classification based techniques rely on availability of accurate labels for various normal classes, which is often not possible.

2) Assigning Label to Each Test Instance – This can become a disadvantage when a meaningful outlier score is desired for the test instances being subject to classification based techniques. However some classification techniques that obtain a probabilistic prediction score from the output of a classifier, can be used to address this issue [58].

IX. CONCLUDING REMARKS AND FUTURE WORK

In this survey we have discussed different ways in which the problem of classification based outlier detection has been formulated in literature, and we attempted to provide an overview of the huge literature on different techniques. For each subcategory of Classification based technique, we could identify a unique assumption regarding the notion of normal data and outliers. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain. We understand that ideally, a comprehensive survey should not only allow a reader to understand the motivation behind using a particular technique, but also provide a comparative analysis of various techniques. But the current research done in an unstructured fashion, without relying on a unified notion of outliers, makes a theoretical understanding of the outlier detection problem a difficult task. A possible future work would be to unify the assumptions made by different techniques regarding the normal and outlier behavior into a statistical or machine learning framework.

X. ACKNOWLEDGMENTS

Dr Shuchita Upadhyaya and Karanjit Singh and thanks the staff of Computer Science and Applications department of Kurukshetra University for their wholehearted support in referencing the study materil. The authors sincerely thank the library staff for the late duty hours at times.

REFERENCES


