An Empirical Comparison of Outlier Detection Methods

Rohan Baxter, Hongxing He, Graham Williams, Simon Hawkins and Lifang Gu

Abstract. Four outlier detection methods are compared using both publicly available smaller statistical datasets and real-life Knowledge Discovery in Databases (KDD) datasets [1]. The smaller datasets provide insight (via visualisations) into the relative strengths and weaknesses of the compared methods. The real-life large datasets test scalability and practicality of application. We are unaware of previous comparisons of outlier detection methods for data mining applications. A methodology for comparing outlier detection methods is developed and we provide performance benchmarks against which new outlier detection methods can be assessed.

(outlier detection, empirical comparison, clustering, neural network, mixture modelling)
An Empirical Comparison of Outlier Detection Methods

Rohan Baxter, Hongxing He, Graham Williams, Simon Hawkins and Lifang Gu
Mathematical and Information Sciences
CSIRO

Abstract. Four outlier detection methods are compared using both publicly available smaller statistical datasets and real-life Knowledge Discovery in Databases (KDD) datasets [1]. The smaller datasets provide insight (via visualisations) into the relative strengths and weaknesses of the compared methods. The real-life large datasets test scalability and practicality of application. We are unaware of previous comparisons of outlier detection methods for data mining applications. A methodology for comparing outlier detection methods is developed and we provide performance benchmarks against which new outlier detection methods can be assessed.
(outlier detection, empirical comparison, clustering, neural network, mixture modelling)

1 Introduction

Studies from the field of statistics have typically considered outliers to be residuals or deviations from a regression or density model of the data:

An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism [5].

We are not aware of previous empirical comparisons of outlier detection methods incorporating both statistical and data mining methods and datasets in the literature. In contrast, the statistical outlier detection literature contains many empirical comparisons between alternative methods. We believe that progress can best be made with some publicly available performance benchmarks against which new outlier detection methods can be assessed. An analogy can be made with the study of classifiers in the machine learning field where benchmark datasets [2] and a standardised comparison methodology [?] allowed the significance of classifier innovation to be properly assessed.

Two parametric and two non-parametric outlier detection methods are compared in this paper. The parametric methods originate from the statistical literature, while the non-parametric methods originate from the data mining literature. A priori, we expect that parametric and non-parametric methods will perform differently on the test datasets. For example, a datum may not lie far
from a very complex model (e.g., a clustering model with many clusters), while it may lie far from a simple model (e.g., a single hyper-ellipsoid cluster model). This leads to the concept of local outliers in the data mining literature [7]. The parametric approach is designed for datasets with a dominating relatively dense convex bulk of data records. The empirical comparisons in this paper show that the parametric approach is still competitive for more complex data. This, perhaps surprisingly, suggests that the parametric approach is sufficient for the real-life datasets examined in this paper.

The outlier detection methods to be compared are reviewed in Section 2. In Section 3, we describe the datasets and experimental design for the comparisons. The test results of the four methods are reported in Section 4. Section 5 summarises the results and contribution of the paper.

2 Outlier Detection Methods

Selection of the four outlier detection methods compared in this paper is based on availability of implementation, our own awareness and familiarity, and our intent to sample from distinctive approaches.

The four chosen outlier detection methods are

- Donoho-Stahel estimator [8]
- Hadi94 [4]
- MML clustering [9]
- RNN [6]

Of course, there are many data mining outlier detection methods not included here [8, ?, ?, ?, ?, 8] and also many omitted statistical outlier methods [10, ?, ?, ?, ?].

Many of the methods not included here are related to the four included methods. Like RNN and MML clustering, many of the proposed data mining outlier detection methods are adapted from clustering methods.

The Donoho-Stahel outlier detection method uses the outlyingness measure computed by the Donoho-Stahel estimator, which is a robust multivariate estimator of location and scatter [8]. It can be characterised as an ‘outlyingness-weighted’ estimate of mean and covariance, which downweights any point that is many robust standard deviations away from the sample in some univariate projection [8].

Hadi94 [4] is a parametric bulk outlier detection method for multivariate data. The method starts with $g = k + 1$ ‘good’ records, where $k$ is the number of dimensions. The good set is increased one point at a time and the $k + 1$ ‘good’ records are selected using a robust estimation method. The mean and covariance matrix of the ‘good records’ are calculated. The Mahalanobis distance is computed for all the data and the $g = g + 1$ closest data are selected as the ‘good records’. This is repeated until the ‘good’ records contain more than half the dataset, or the Mahalanobis distance of the remaining records is higher than a predefined cut-off value.
We use the mixture-model clustering algorithm where models are scored using MML inductive inference [9]. The cost of transmitting each datum according the best mixture model is measured in nits (1.6 bits = 1 nit). We rank the data in order of highest message length cost to lowest message length cost. The high cost data are ‘surprising’ according to the model and so are considered as outliers by this method.

The replicator neural network (RNN) method uses a multi-layer perceptron neural network with three hidden layers and the same number of output nodes and input nodes. The input data are used as the targetted output data while training the RNN model. A measure of outlyingness of individuals is then developed as the reconstruction error of a datum using the trained RNN model.

## 3 Experimental Design

We now describe the test datasets and their characteristics.

The statistical outlier detection literature has considered three qualitative types of outliers. First, *cluster outliers* occur in small low variance clusters. The ‘low variance’ is relative to the variance of the bulk of the data. *Radial outliers* occur in a plane out from the major axis of the bulk of the data. If the bulk of data occurs in an elongated ellipse then radial outliers will lie on the major axis of that ellipse but separated from and less densely than the bulk of data. *Scattered outliers* occur randomly scattered about the bulk of data. The included datasets in Section 4 contain different combinations of outliers from these three outlier categories. Datasets will have different proportions of outliers in the data. The proportion of outliers is called the *contamination level*.

The datasets from the statistical literature used in this paper are shown in Table 1. A description of the original source of the datasets and the datasets themselves are found in [11]. These datasets are used throughout the statistical outlier detection literature. The KDD datasets used are listed in Table 2. Note that all the datasets are publically available [1, 11].

Most other data mining papers use simulation studies rather than real world datasets. We provide a visualisation of some of the datasets in Section 4 using the visualisation tool *xgobi* [12].

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Number of Records (n)</th>
<th>Number of Dimensions (k)</th>
<th>Outliers</th>
<th>%</th>
<th>Outlier Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBK</td>
<td>75</td>
<td>4</td>
<td>14</td>
<td>21</td>
<td>small cluster, also some scattered</td>
</tr>
<tr>
<td>Wood</td>
<td>20</td>
<td>6</td>
<td>4</td>
<td>20</td>
<td>radial (on axis), and in small cluster</td>
</tr>
<tr>
<td>Milk</td>
<td>85</td>
<td>8</td>
<td>17</td>
<td>18</td>
<td>radial, also some scattered off main axis</td>
</tr>
<tr>
<td>Hertzsprung</td>
<td>47</td>
<td>2</td>
<td>7</td>
<td>22</td>
<td>some scattered, some in a cluster, scattered</td>
</tr>
<tr>
<td>Stackloss</td>
<td>21</td>
<td>4</td>
<td>4</td>
<td>25</td>
<td>scattered</td>
</tr>
</tbody>
</table>

Table 1. Statistical Outlier Detection Test Datasets [11]
<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Number of Records ($n$)</th>
<th>Number of Dimensions ($k$)</th>
<th>$\frac{n}{k}$</th>
<th>%</th>
<th>Outlier Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrusion-http</td>
<td>367497</td>
<td>3</td>
<td>2211</td>
<td>4</td>
<td>200K small separate cluster.</td>
</tr>
<tr>
<td>Intrusion-smtp</td>
<td>95156</td>
<td>3</td>
<td>30</td>
<td>0.3</td>
<td>30K scattered outlying but also some between two elongated different sized clusters.</td>
</tr>
<tr>
<td>Intrusion-ftp-data</td>
<td>30464</td>
<td>3</td>
<td>722</td>
<td>2</td>
<td>10K outlying cluster. also some scattered.</td>
</tr>
<tr>
<td>Intrusion-other</td>
<td>5858</td>
<td>3</td>
<td>98</td>
<td>2</td>
<td>2K 2 clusters, both overlapping with non-intrusion data.</td>
</tr>
<tr>
<td>Intrusion-ftp</td>
<td>4091</td>
<td>3</td>
<td>316</td>
<td>8</td>
<td>1K scattered outlying and a cluster.</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>683</td>
<td>9</td>
<td>239</td>
<td>35</td>
<td>76 scattered.</td>
</tr>
</tbody>
</table>

Table 2. KDD Outlier Detection Test Datasets [1]

4 Experimental Results

Due to space limitations, only the HBK and Wood results are given here. The full paper has the results for the remaining three statistical datasets.

4.1 HBK

Figure 1 is a visualisation of the outliers for the HBK dataset. The outlier records are well-separated from the bulk of the data and so we may expect the outlier detection methods to easily distinguish the outliers. The results from the outlier detection methods for the HBK dataset are summarised in Table 3. Donoho-Stahel and Hadi94 rank the 14 outliers in the top 14 places and the distance measures dramatically distinguish the outliers from the remainder of the records. MML clustering does less well. It identifies the scattered outliers but the outlier records occurring in a compact cluster are not ranked as outliers. This is because their compact occurrence leads to a small description length. RNN has the 14 outliers in the top 16 places and has placed all the true outliers in a single cluster.

Fig. 1. Visualisation of HBK dataset using xgobi. The 14 outlier records are represented by the small filled-in squares.
Table 3. Top 20 outliers for Donoho-Stahel, Hadi94, MML Clustering and RNN on the HBK dataset. `True` outliers are in bold. There are 14 outliers indexed 1..14.

4.2 Wood Data

Figure 2 is a visualisation of the outliers for the Wood dataset. The dataset is from [11], where it is noted that the four outliers (observations 4, 6, 8, 19) were created by modifying the original dataset. The results from the outlier detection methods for the Wood dataset are summarised in Table 4. Donoho-Stahel identifies the four outlier records, while Hadi94, RNN and MML do not.

The difference between Donoho-Stahel and Hadi94 is interesting and can be explained by their different estimates of scatter (or covariance). Donoho-Stahel’s estimate of covariance is more compact (leading to a smaller ellipsoid around the estimated data centre). This result empirically suggests Donoho-Stahel’s improved robustness with high dimensional datasets relative to Hadi94.

MML clustering has considerable difficulty in identifying the outliers according to description length and it ranks the true outliers last! The cluster membership column allows an interpretation of what has happened. MML clustering puts the outlier records in their own low variance cluster and so the records are described easily at low information cost. Identifying outliers by rank using data
Table 4. Top 20 outliers for Donoho-Stahel, Hadi94, MML Clustering and RNN on the Wood dataset. ‘True’ outliers are in bold. The 4 true outliers are 4, 6, 8, 19.

description length with MML clustering does not work for low variance cluster outliers.

For RNN, the cluster membership column again allows an interpretation of what has happened. Most of the data belong to cluster 0, while the outliers belong in various other clusters. Similarly to MML clustering, the outliers can be identified by interpreting the clusters.

4.3 Network Intrusion Detection

This dataset comes from the 1999 KDD Cup network intrusion detection competition [3]. We follow the experimental technique employed in [14, 6, 13]. Attack events are treated as outliers from normal events.

We select four of the 41 original attributes (service, duration, src_bytes, dst_bytes). These attributes are thought to be the most important features [14]. Service is a categorical feature while the other three are continuous features.

The original dataset contained 4,898,431 data records, including 3,925,651 attacks (80.1%). This high rate is too large for attacks to be considered outliers. Therefore, following [14] we produced a subset consisting of 703,066 data records including 3,377 attacks (0.48%). The subset consists of those records with logged_in being positive.

The dataset was then divided into five subsets according the five values of the service variable. The results for each subset are now discussed.

A visualisation of the other dataset (not shown here for space reasons) shows that half the attacks are occurring in a distinct outlying cluster, while the other half occur embedded amongst normal events. Figure 3 summarises the results for the four methods on the other dataset. RNN finds the first 40 outliers long before any of the other methods. All the methods need to see more than 60% of
the observations before including 80 of the total (98) outliers in their rankings. This suggests there is low separation between the bulk of the data and the outliers.

A visualisation of the http dataset (not shown here) shows that the intrusions occur in a small cluster separate from the bulk of the data. Figure 4 summarises the results for the four methods on the KDD other data. The performance of Donoho-Stahel, Hadi94 and RNN cannot be distinguished. MML Clustering needs to see an extra 10% of the data before including all the intrusions.

A visualisation of the smtp dataset (not shown here) shows that most intrusions occur very separately from the bulk of the data. Figure 5 summarises the results for the four methods on the smtp dataset. The performance of Donoho-Stahel, Hadi94 and MML trend very similarly while RNN needs to see nearly all of the data to identify the last 40% of the intrusions.

A visualisation of the ftp dataset shows that most intrusions appear quite separate from the bulk of the data. Figure 7 summarises the results for the four methods on the ftp dataset. The performance of Donoho-Stahel and Hadi94 trend very similarly. RNN needs to see ≈ 20% more of the data to identify most of the intrusions. MML clustering does not do much better than random in ranking the intrusions above normal events. Only some intrusions are scattered, while the remainder lie in clusters of a similar shape to the normal events.

A visualisation of the ftp-data shows that most intrusions appear quite separate from the bulk of the data. Figure 6 summarises the results for the four methods on the ftp-data dataset. Donoho-Stahel performs the best. RNN needs to see 20% more of the data. Hadi94 needs to see another 20% more. The MML curve is below the y = x curve (which would arise if a method randomly ranked
**Fig. 4.** Ratio of detected intrusions found by the Donoho-Stahel, Hadi94, MML Clustering and RNN methods on the KDD http data

**Fig. 5.** Ratio of detected intrusions found by the Donoho-Stahel, Hadi94, MML Clustering and RNN methods on the KDD smtp data
Fig. 6. Ratio of detected intrusions found by the Donoho-Stahel, Hadi94, MML Clustering and RNN methods on the KDD *ftp-data* data, indicating that the intrusions have been placed in low variance clusters requiring small description lengths.

Fig. 7. Ratio of detected intrusions found by the Donoho-Stahel, Hadi94, MML Clustering and RNN methods on the KDD *ftp* data

### 4.4 Wisconsin Breast Cancer Dataset

Overall, all the methods except Donoho-Stahel, have little difficulty identifying the outliers in this dataset. So we sample the original dataset to generate datasets with differing contamination levels (number of malignant observations) ranging from 8.07% to 35%. We now investigate the performance of the methods with
differing contamination levels. Figure 8 is a visualisation of the outliers for this dataset.

![Visualisation of Breast Cancer dataset using xgobi. Malignant data are shown as grey crosses.](image)

**Fig. 8.** Visualisation of Breast Cancer dataset using xgobi. Malignant data are shown as grey crosses.

Figure 9 shows the coverage of outliers by the Hadi94 method versus % observations for various level of contamination. The performance of Hadi94 degrades as the level of contamination increases, as one would expect. The results for the MML clustering method and the RNN method track the Hadi94 method closely and are not shown here. The Donoho-Stahel method does not do any better than a random ranking of the outlyingness of the data. Investigating further we find that the robust estimate of location and scatter is quite different to that of Hadi94 and obviously less successful.

![Hadi94 outlier performance as outlier contamination varies](image)

**Fig. 9.** Hadi94 outlier performance as outlier contamination varies
5 Discussion and Conclusion

This paper makes contributions in the areas of:

– Understanding and categorising some publically available benchmark datasets for testing outlier detection algorithms.

– Comparing the performance of four different outlier detection methods; two parametric methods from the statistical literature and two non-parametric methods from the machine learning/data mining literature.

– Using outlier categories: cluster, radial and scattered and contamination levels to characterise the difficulty of the outlier detection task for large KDD datasets (as well as the usual statistical test datasets).

We conclude that the statistical outlier detection method, Hadi94, scales well and performs well on large and complex datasets. The Donoho-Stahel method matches the performance of the Hadi94 method in general except for the breast cancer dataset. Since this dataset is relatively large in size ($n = 664$) and dimension ($k = 9$), this suggests that the Donoho-Stahel method does not handle this combination of large $n$ and large $k$ well. We plan to investigate whether the Donoho-Stahel method’s performance can be improved by using different heuristics for the number of sub-samples used by the sub-sampling algorithm (described in Section 2).

The RNN method’s performance varied for both the small and large datasets. That it performed well on the small datasets at all is interesting because neural network methods often have difficulty with smaller datasets. Its performance appears to degrade with datasets containing radial outliers and so it is not recommended for this type of dataset. RNN performed the best overall on the KDD intrusion dataset.

The MML clustering method works well for scattered outliers. For cluster outliers, the user needs to look for small population clusters and then treat them as outliers, rather than just use the ranked description length method (as was used here).

Outlier detection is, like clustering, an unsupervised classification problem where simple performance criteria based on accuracy, precision or recall do not easily apply. In this paper we have presented results as a ranking based on each method’s outlyingness measure.

References


