A Comparative Study of RNN for Outlier Detection in Data Mining

Graham Williams Rohan Baxter Hongxing He Simon Hawkins
Lifang Gu
CSIRO Enterprise Data Mining
GPO Box 664, Canberra, ACT 2601, Australia
Firstname.Lastname@csiro.au
http://datamining.csiro.au

Abstract

We have proposed replicator neural networks (RNNs) for outlier detection [8]. Here we compare RNN for outlier detection with three other methods using both publicly available statistical datasets (generally small) and data mining datasets (generally much larger and generally real data). The smaller datasets provide insights into the relative strengths and weaknesses of RNNs. The larger datasets particularly test scalability and practicality of application.

1. Introduction

Outlier detection has regained considerable interest in data mining with the realisation that they can be the key discovery from very large databases [5, 4, 16]. Indeed, for many applications the discovery of outliers leads to more interesting and useful results than the discovery of inliers. The classic example is fraud detection but in customer relationship management (CRM) and other consumer databases outliers are often the most profitable group of customers.

In this paper we apply replicator neural networks (RNNs) to outlier detection [8]. RNNs have a flexible, non-parametric representation of clusters, which, in principle, should make them a useful, powerful outlier detection method. RNNs are compared with two parametric (from the statistical literature) methods and one non-parametric outlier detection method (from the data mining literature).

In Section 2 we review RNN outlier detection and briefly summarise the three alternative methods in Section 3. In Section 4 we describe the datasets and experimental design for performing the comparisons. Results are reported in Section 5, and Section 6 summarises the results and the contribution of this paper.

2. RNNs for Outlier Detection

RNNs for outlier detection [8]. employ feed-forward multi-layer neural networks with three hidden layers sandwiched between the input and output layers which have n units each, corresponding to the n features of the training data. The number of units in the three hidden layers are chosen experimentally to minimise the average reconstruction error across all training patterns. The neural network input variables are *also* the output variables so that the RNN forms an implicit, compressed model of the data during training—the RNN attempts to reproduce the input patterns in the output. A measure of outlyingness of individuals is then developed as the reconstruction error of individual data points.

A particular innovation is the activation function used for the middle hidden layer [8]. Instead of the usual sigmoid activation function for this layer (layer 3) a staircase-like function with parameters N (number of steps or activation levels) and a_3 (transition rate from one level to the next) are employed. As a_3 increases the function approaches a true step function, as shown in Figure 1.

This activation function has the effect of dividing continuously distributed data points into a number of discrete valued vectors, providing the data compression that RNNs are known for. For outlier detection the mapping to discrete categories naturally places the data points into a number of clusters. For scalability the RNN is trained with a smaller training set and then applied to all of the data to evaluate their outlyingness.

3. Other Outlier Detection Methods

Selection of the outlier detection methods used in this paper for comparison is based on availability and our intent to sample from distinctive approaches. The three chosen methods are: the Donoho-Stahel estimator [10]; Hadi94 [6]; and MML clustering [12].

0-7695-1754-4/02 $17.00 © 2002 IEEE

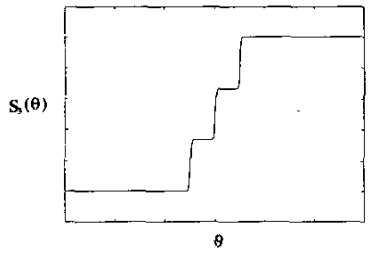

Figure 1. A representation of the activation function for units in the middle hidden layer of the RNN. N is 4 and a_3 is 100 [8].

Of course, there are many data mining outlier detection methods not included here [9] and also many omitted statistical outlier methods [13, 11]. Many of these methods are related to the three included methods and RNNs, often being adapted from clustering methods in one way or another and a full appraisal of this is worth a separate paper.

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 [10]. It can be characterised as an 'outlyingness-weighted' estimate of mean and covariance, which down-weights any point that is many robust standard deviations away from the sample in some univariate projection.

Hadi94 [6] is a parametric bulk outlier detection method for multivariate data. The method starts with $g_0 = k + 1$ '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 $g_n = g_{n-1} + 1$ 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 threshold value.

For the data mining outlier detector we use the mixture-model clustering algorithm where models are scored using MML inductive inference [12]. The cost of transmitting each datum according to the best mixture model is measured in nits (1.6bits = 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.

4. Experimental Design

Each outlier detection method has a bias toward its own implicit or explicit model of outlier determination. A variety of datasets are needed to explore the differing biases and to develop an understanding of the appropriateness of each method for particular characteristics of data sets.

The statistical outlier detection literature considers three qualitative outlier categories: *cluster outliers* occur in small low variance clusters (where 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 separated from and less densely packed than the bulk of data); *scattered outliers* occur randomly scattered about the bulk of data.

The datasets contain different combinations of outliers from these categories. We refer to the proportion of outliers in a data set as the *contamination level* of the data set and include datasets that exhibit different proportions of outliers. Statistical literature typically considers contamination levels of up to 40% whereas data mining literature typically considers contamination levels of much less ($< 4\%$), typical of the types of outliers in, for example, fraud (where it is even as low as 1% or less). Identifying 40% of a very large dataset as outliers is unlikely to provide useful insights into these rare, yet very significant, groups.

The datasets from the statistical literature and the data mining literature used in this paper are listed in Table 1.

Dataset	n	k	o	%	$\frac{n}{k}$	Description
HBK	75	4	14	21	19	Small cluster, some scattered.
Wood	20	6	4	20	3	Radial, small cluster.
Cancer	683	9	239	35	76	Scattered.
http	567497	3	2211	0.4	200K	Separate cluster.
smtp	95156	3	30	0.03	30K	Scattered, outlying.
ftp-data	30464	3	722	2	10K	Outlying cluster, some scattered.
other	5858	3	98	2	2K	Two clusters.
ftp	4091	3	316	8	1K	Scattered outlying and cluster.

Table 1. Statistical [14] and data mining outlier detection test datasets [2, 17]. Number of records = n, dimensions = k, outliers = o.

We can observe that outliers from the statistical datasets arise from measurement errors or data-entry errors, while the outliers in the selected data mining datasets are semantically distinct categories. Thus, for example, the breast cancer data has non-malignant and malignant measurements and the malignant measurements are viewed as outliers. The intrusion dataset identifies successful Internet intrusions. Intrusions are identified as exploiting one of the possible vulnerabilities, such as *http* and *ftp*. Successful intrusions are considered as outliers in these datasets.

It would have been preferable to include more data mining datasets for assessing outlier detection. The KDD intrusion dataset is included because it is publicly available and has been used previously in the data mining literature [17].

Knorr *et al.* [10] use NHL player statistics but refer only to a web site publishing these statistics, not the actual dataset used. Most other data mining papers use simulation studies rather than real world datasets.[1]

5. Experimental Results

We discuss here results from a selection of the test datasets we have employed to evaluate the performance of the RNN approach to outlier detection. Detailed results are available from the authors [15].

5.1. HBK

The HBK dataset is artificially constructed [7] with 14 outliers. Regression approaches tend to find only the first 10 as outliers. Data points 1-10 are "bad leverage" points—they lie far away from the centre of the good points and from the regression plane. Data points 11-14 are good leverage points—although they lie far away from the bulk of the data they still lie close to the regression plane.

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, identifying the scattered outliers but missing the outlier records occurring in a compact cluster. 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.

5.2. Wood Data

The Wood dataset consists of 20 observations [3] with data points 4, 6, 8, and 19 being outliers [14]. The outliers are said not to be easily identifiable by observation [14].

We found that Donoho-Stahel clearly identifies the four outlier records, while Hadi94, RNN and MML all struggle to identify them. 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! 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 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 a single cluster, while the outliers belong to various other clusters. Similarly to MML clustering, the outliers can, however, be identified by interpreting the clusters.

[1] A collection of data sets for outlier detection are available from `http://datamining.csiro.au/outliers`.

5.3. Wisconsin Breast Cancer Dataset

Our initial exploration of this dataset found that all the methods except Donoho-Stahel have little difficulty identifying the outliers. So we sampled the original dataset to generate datasets with differing contamination levels (number of malignant observations) ranging from 8.07% to 35% to investigate the performance of the methods with differing contamination levels. We found that 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. 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.

5.4. Network Intrusion Detection

The network intrusion dataset comes from the 1999 KDD Cup network intrusion detection competition [1]. We follow the experimental technique employed in [17] to construct suitable datasets for outlier detection and to rank all data points with an outlier measure.

The dataset is divided into five subsets according to the five values of the *service* variable (*other*, *http*, *smtp*, *ftp*, and *ftp-data*). The aim is to identify intrusions within each of the categories by identifying outliers.

For the *other* dataset we observed that half the attacks are occurring in a distinct outlying cluster, while the other half are embedded among normal events. For the *http* dataset intrusions occur in a small cluster separated from the bulk of the data. For the *smtp*, *ftp*, and *ftp-data* datasets most intrusions also appear quite separated from the bulk of the data.

For 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. For the *http* dataset 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. For the *smtp* dataset the performances 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. For the *ftp* dataset the performances of Donoho-Stahel and Hadi94 trend very similarly. RNN needs to see $\approx$ 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. Finally, for 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. MML puts the intrusions in low variance clusters, where they have the shortest description length, and does not identify any scattered outliers with a high description length.

6. Discussion and Conclusion

The main contributions of this paper are:

- Empirical evaluation of the RNN approach for outlier detection;
- Using outlier categories: *cluster*, *radial* and *scattered* and contamination levels to characterise the difficulty of the outlier detection task for large data mining datasets (as well as the usual statistical test datasets).
- Understanding and categorising some publicly available benchmark datasets for testing outlier detection algorithms using outlier categories and contamination levels.
- Comparing the performance of three different outlier detection methods from the statistical and data mining literatures with RNN.

We conclude that the statistical outlier detection methods, Hadi94 and Donoho-Stahel, scale well and perform well on large and complex datasets. However they are parametric methods and lack the flexibility of non-parametric methods such as RNN and MML. MML clustering works well for *scattered* outliers and places *radial* and *cluster* outliers into their own clusters, requiring interpretation of clusters rather than a straight ranking based on an outlyingness measure.

The RNN method performed satisfactorily for both small and large datasets. It was of interest that it performed well on the small datasets since neural network methods often have difficulty with such 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.

References

[1] 1999 KDD Cup competition. `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`.

[2] S. D. Bay. The UCI KDD repository, 1999. `http://kdd.ics.uci.edu`.

[3] N. R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley and Sons, New York, 1966.

[4] W. DuMouchel and M. Schonlau. A fast computer intrusion detection algorithm based on hypothesis testing of command transition probabilities. In *Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD98)*, pages 189–193, 1998.

[5] T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

[6] A. Hadi. A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, B*, 56(2), 1994.

[7] D. M. Hawkins, D. Bradu, and G. V. Kass. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, 26:197–208, 1984.

[8] S. Hawkins, H. X. He, G. J. Williams, and R. A. Baxter. Outlier detection using replicator neural networks. In *Proc. of the Fifth Int. Conf. and Data Warehousing and Knowledge Discovery (DaWaK02)*, 2002.

[9] E. Knorr, R. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *Very Large Data Bases*, 8(3–4):237–253, 2000.

[10] E. M. Knorr, R. T. Ng, and R. H. Zamar. Robust space transformations for distance-based operations. In *Proc. of the 7th Int. Conf. on Knowledge Discovery and Data Mining (KDD01)*, pages 126–135, 2001.

[11] A. S. Kosinksi. A procedure for the detection of multivariate outliers. *Com. Stat. & Data Analysis*, 29, 1999.

[12] J. J. Oliver, R. A. Baxter, and C. S. Wallace. Unsupervised Learning using MML. In *Proc. of the Thirteenth Int. Conf. (ICML 96)*, pages 364–372. Morgan Kaufmann Publishers, San Francisco, CA, 1996.

[13] D. E. Rocke and D. L. Woodruff. Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91:1047–1061, 1996.

[14] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York, 1987.

[15] G. Williams, R. Baxter, H. He, S. Hawkins, and L. Gu. A comparative study of RNN for outlier detection in data mining. Technical Report 02/102, CSIRO Mathematical and Information Sciences, 2002. http://datamining.csiro.au/papers/tr02102.pdf.

[16] G. J. Williams and Z. Huang. Mining the knowledge mine: The hot spots methodology for mining large real world databases. In A. Sattar, editor, *Advanced Topics in Artificial Intelligence*, volume 1342 of *Lecture Notes in Artificial Intelligenvce*, pages 340–348. Springer, 1997.

[17] K. Yamanishi, J. Takeuchi, G. J. Williams, and P. W. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithm. In *Proc. of the 6th Int. Conf. on Knowledge Discovery and Data Mining (KDD00)*, pages 320–324, 2000.