1 Retrieval data sets

COREL

The COREL image database [12] of 10000 images. There are 100 categories and each category has 100 images. (We are not able to provide the data due to copyright issue.)

MATLAB data format

Each MATLAB data file is composed of

- $Data$ — $n \times d$ data matrix, where n is the number of images and d is the dimension;
- $Labels$ — $n \times k$ label matrix, where k is the number of categories; $Labels(i, j) = 1$ if image(i) in category j; otherwise, $Labels(i, j) = -1$;

2 Regression data sets

tic

The Insurance Company Benchmark (COIL 2000) data which is available at UCI repository [1].

wine

The White and Red Wine Quality data which is available at UCI repository [2].

quake

The earthquake data [9] which is converted from WEKA quake data in "datasets-numeric.jar" (available at [3]).

concrete

The Concrete Compressive Strength data which is available at UCI repository [4].

MATLAB data format

Each MATLAB data file is composed of

- $Data$ — $n \times d$ data matrix, where n is the number of instances and d is the dimension;
- $Labels$ — $n \times 1$ vector of regression values;

3 Anomaly detection data sets

Http and Smtp

The two data sets are picked from KDD Cup 1999 data, which is available at UCI repository [5]. The original KDD Cup 1999 training data contain 41 attributes, however, they are reduced to 4 attributes (service, duration, src_bytes, dst_bytes) as these attributes are regarded as the most basic attributes. Using the 'service' attribute, the data is divided into {http, smtp, ftp, ftp_data, others} subsets. Other attributes are transformed by $y = log(x+0.1)$. The original data set has 3,925,651 attacks (80.1%) out of 4,898,431 records. A smaller set is forged by having only 3,377 attacks (0.35%) of 976,157 records. These subsets are first used by [11] and subsequently used by [10].

In our experiment, we use the largest two subsets, they are: Http (567,497 records) and Smtp (95,156 records). The anomalies ratios are 0.4% for http and 0.03% for smtp.

Forest

The Covertype data which is available at UCI repository [6]. In our experiments, instances from class 2 are considered as normal points and instances from class 4 are anomalies. The anomalies ratios is 0.9%. Instances from the other classes are omitted.

Mulcross

The data is generated from a synthetic data generator Mulcross [8] and available at [7]. Mulcross generates a multi-variate normal distribution with a selectable number of anomaly clusters. In our experiments, the basic setting for Mulcross is as following: contamination ratio = 10% (number of anomalies over the total number of points), distance factor = 2 (distance between the center of normal cluster and anomaly clusters), and number of anomaly clusters = 2.

Shuttle

The Statlog (Shuttle) data which is available at UCI repository [6]. In our experiments, instances from class 1 are considered as normal points and instances from class 2, 3, 5, 6, 7 are anomalies. The anomalies ratios is 7.15%. Instances from the other classes are omitted.

MATLAB data format

Each MATLAB data file is composed of

- *Data* — $n \times d$ data matrix, where n is the number of instances and d is the dimension;

- $ADLabels$ — $n \times 1$ vector of anomaly labels; $ADLabels(i) = 1$ if instance(i) is anomaly; otherwise, $ADLabels(i) = 0$;

References

[1] http://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+(COIL+2000).

[2] http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

[3] http://www.cs.waikato.ac.nz/ml/weka.

[4] http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength.

[5] http://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data.

[6] http://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle).

[7] http://personal.gscit.monash.edu.au/ kmting/Mass.

[8] David M. Rocke and David L. Woodruff. Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061, 1996.

[9] Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer-Verlag, 1996.

[10] Graham Williams, Rohan Baxter, Hongxing He, Simon Hawkins, and Lifang Gu. A comparative study of rnn for outlier detection in data mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, page 709, Washington, DC, USA, 2002. IEEE Computer Society.

[11] Kenji Yamanishi, Jun-Ichi Takeuchi, Graham Williams, and Peter Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 320–324, New York, NY, USA, 2000. ACM Press.

[12] Zhi-Hua Zhou, Ke-Jia Chen, and Hong-Bin Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 24(2):219–244, 2006.