Mach Learn (2016) 105:301–304
DOI 10.1007/s10994-016-5566-8

Commentary: a decomposition of the outlier detection problem into a set of supervised learning problems

Ye Zhu[1] · **Kai Ming Ting**[2]

Received: 29 February 2016 / Accepted: 9 April 2016 / Published online: 29 April 2016

Abstract This article discusses the material in relation to *i*Forest (Liu et al. in ACM Trans Knowl Discov Data 6(1):3, 2012) reported in a recent Machine Learning Journal paper by Paulheim and Meusel (Mach Learn 100(2–3):509–531, 2015). It presents an empirical comparison result of *i*Forest using the default parameter settings suggested by its creator (Liu et al. 2012) and *i*Forest using the settings employed by Paulheim and Meusel (2015). This comparison has an impact on the conclusion made by Paulheim and Meusel (2015).

Keywords Anomaly detection · Outlier detection · Isolation forest

1 Background

Paulheim and Meusel (2015) (referred to as PM hereafter) proposed an outlier detection method based on supervised regression learning, named attribute-wise learning for scoring outliers (ALSO). For a dataset of d attributes, this method builds d linear regression models, where each linear regression model predicts the value of one of the d attributes using all other attributes. During the training process, it calculates deviations between predicted values and actual values. The final outlier score is a weighted average of these deviations over all attributes of the dataset, where the weight for an attribute is the root relative squared error of the regression. This method can handle data with high dimensionality and is tolerant to irrelevant attributes due to its attribute weighting.

Editor: Peter Flach.

✉ Ye Zhu
yale.zhu@monash.edu

[1] Faculty of Information Technology, Monash University, Melbourne, VIC, Australia

[2] School of Engineering and Information Technology, Federation University, Churchill, VIC, Australia

PM compared ALSO with 10 contenders on real-world datasets from the UCI machine learning repository (Lichman 2013), and reported that ALSO using M5' (Quinlan 1992) to build the regression models was significantly better than most of the 10 contenders.

*i*Forest (isolation Forest), one of the closest contenders reported by PM, is of particular interest because the parameter settings used in PM's experiments are not the default settings as suggested in the *i*Forest paper (Liu et al. 2012). (Details of the two settings are given in the Discussion section, and a brief description of *i*Forest is provided in the Appendix for ease of reference.)

This article presents a comparison of these two different settings of *i*Forest to examine a claim made by PM, and provides a discussion on the material in relation to *i*Forest reported by PM.

2 Comparison of two settings of *i*Forest as suggested by Paulheim and Meusel (2015) and Liu et al. (2012)

We have tested *i*Forest using the default parameter settings [suggested by Liu et al. (2012)] on the 12 datasets provided by PM. For each dataset, we report the average AUC (Area under ROC curve) results over 10 trials with different seeds for randomisation.

Table 1 presents the results of the two settings of *i*Forest: one provided by Paulheim (private communication) and the other obtained using the default settings. These results show that there are 9 wins and 3 losses out of the 12 datasets, in favour of *i*Forest with the default settings; and many of the wins are in large margins. The average AUC over the 12 datasets is 0.845; whereas the average AUC result from PM is 0.781, which is significantly less than that using the default settings. The results of two significance tests are given below.

Table 1 AUC of *i*Forest: The result from PM versus the result using the default settings (Liu et al. 2012)

Dataset	*i*Forest PM's parameter settings (ψ used not stated; $t = 30$; $hlim = 1$)	*i*Forest Default parameter settings ($\psi = 256$; $t = 100$; $hlim = \psi - 1$)
Shuttle	0.992	**0.999**
Satellite	0.916	**0.956**
Ionosphere	**0.947**	0.938
Breast Cancer	0.803	**0.980**
Glass	0.923	**0.947**
Seismic	0.718	**0.727**
Parkinson	**0.396**	0.368
Concrete	**0.694**	0.675
Wine	0.572	**0.667**
Energy	0.628	**0.970**
CCPP	0.906	**0.946**
Housing	0.882	**0.963**
average	0.781	**0.845**
Win/Loss	—	**9/3**

The bold number indicates the better performer. The details of both settings can be found in the Discussion section

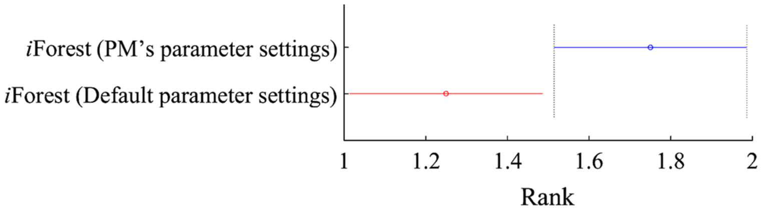

Fig. 1 Friedman test result with the post-hoc Nemenyi test at $p = 0.10$. The difference between the two settings of *i*Forest is significant because their critical differences (*horizontal bars*) do not overlap

We have conducted a Friedman test with the post-hoc Nemenyi test (Demšar 2006) to examine whether the difference between the two settings of *i*Forest is significant. Figure 1 shows the average rank and critical difference of each setting of *i*Forest. This test shows that *i*Forest with the default parameter settings performs significantly better than *i*Forest with the settings used by PM.

PM used one-sided pared t test and claimed that ALSO(M5') is significantly better than *i*Forest with the settings used by PM with $p < 0.05$. Using the same significance test, we found that *i*Forest with the default parameter settings performs significantly better than *i*Forest with the settings used by PM since the p value we got is 0.03.

On the 12 datasets, *i*Forest with the default parameter settings achieved an average AUC of 0.845 which is comparable to the best result of 0.854, achieved by ALSO(M5') reported by PM.

In summary, the difference between *i*Forest and ALSO(M5') is considerably smaller than that reported by PM, had the default settings suggested by Liu et al. (2012) been used in PM's experiments.

3 Discussion

In Section 4.2 of their paper, PM state that "As reported in Liu et al. (2012), isolation forests provide stable results if at least 30 trees are learned, and the best results are achieved with a height limit of 1, so we use those values." These are not the default settings suggested by Liu et al. (2012); and the most important parameter for *i*Forest, i.e., subsampling size, is not mentioned at all by PM.

The default settings for the three parameters, suggested by Liu et al. (2012), are given as follows:

> "*Subsampling Size*. .... we also find that setting ψ to 2^8 or 256 generally is enough..."
> "Number of trees t controls the ensemble size. We find that path lengths usually converge well before $t = 100$. Unless otherwise specified, we use $t = 100$ as the default value in our experiment."
> "In the normal usage of *i*Forest, the default value of evaluation height limit is set to maximum, that is, $\psi - 1$,..."

Although Liu et al. (2012) have made a suggestion to use the height limit of 1 in Section 5.3, that particular suggestion was made to illustrate a special case using an artificial dataset: " we find that setting a lower evaluation height limit is effective in handling dense anomaly clusters. *i*Forest obtains its best performance using $hlim = 1$. It is because *i*Forest uses the coarsest granularity to detect clustered anomalies."

In summary, PM have used settings other than the default settings for *i*Forest suggested by Liu et al. (2012) in their experiments. The reported *i*Forest result by PM is significantly worse than that obtained using the default parameter settings suggested by Liu et al. (2012), as shown in the comparison result presented in the last section.

Appendix: Isolation forest (*i*Forest)

*i*Forest (Liu et al. 2012) employs a completely random isolation mechanism to isolate every instance in the given training set. This is done efficiently by random axis-parallel partitioning (without using a test selection criterion) of the data space in a tree structure until every instance is isolated. An *i*Forest consists of t trees, and each tree is built using a subsample randomly selected from the given dataset. The anomaly score of an instance $\mathbf{x}$ is measured as the average path length over t trees as follows:

$$s(\mathbf{x}) = \frac{1}{t}\sum_{i=1}^{t} \ell_i(\mathbf{x}) \tag{1}$$

where $\ell_i(\mathbf{x})$ is the path length of $\mathbf{x}$ in tree i.

The intuition is that anomalies are more susceptible to isolation. *i*Forest identifies anomalies as instances having the shortest average path lengths in a dataset.

*i*Forest is one of the fastest anomaly detectors and performs competitive to the state-of-the-art in terms of detection accuracy (Bandaragoda 2015).

Emmott et al. (2013), whom have used the default parameter settings of *i*Forest suggested by Liu et al. (2012) in their independent evaluation, reported that *i*Forest outperformed One-Class SVM algorithm (Schölkopf et al. 2001), Support Vector Data Description (Tax and Duin 2004) and Local Outlier Factor (Breunig et al. 2000).

Acknowledgments The anonymous reviewers have provided many helpful comments to improve the presentation of this article.

References

Bandaragoda, T. R. (2015). *Isolation based anomaly detection: A re-examination*. Master's thesis, Faculty of Information Technology, Monash University.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM Sigmod Record*, *29*, 93–104.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1–30.

Emmott, A. F., Das, S., Dietterich, T., Fern, A., & Wong, W. K. (2013). Systematic construction of anomaly detection benchmarks from real data. In: *ACM SIGKDD workshop on outlier detection and description*, pp. 16–21.

Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(1), 3.

Paulheim, H., & Meusel, R. (2015). A decomposition of the outlier detection problem into a set of supervised learning problems. *Machine Learning*, *100*(2–3), 509–531.

Quinlan, J. R. (1992). Learning with continuous classes. In *The fifth Australian joint conference on artificial intelligence*, Singapore, pp. 343–348.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*(7), 1443–1471.

Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine Learning*, *54*(1), 45–66.