Extraction of Outliers from Imbalanced Sets

Pavel Škrabánek[1] and Natália Martínková[2,3]

[1] Faculty of Electrical Engineering and Informatics, University of Pardubice,
Studentská 95, 532 10 Pardubice, Czech Republic,
pavel.skrabanek@upce.cz

[2] Institute of Vertebrate Biology, Czech Academy of Sciences,
Květná 8, 603 65 Brno, Czech Republic,
martinkova@ivb.cz

[3] Institute of Biostatistics and Analyses, Masaryk University,
Kamenice 3, 625 00 Brno, Czech Republic

Abstract. In this paper, we presented an outlier detection method, designed for small datasets, such as datasets in animal group behaviour research. The method was aimed at detection of global outliers in unlabelled datasets where inliers form one predominant cluster and the outliers are at distances from the centre of the cluster. Simultaneously, the number of inliers was much higher than the number of outliers. The extraction of exceptional observations (EEO) method was based on the Mahalanobis distance with one tuning parameter. We proposed a visualization method, which allows expert estimation of the tuning parameter value. The method was tested and evaluated on 44 datasets. Excellent results, fully comparable with other methods, were obtained on datasets satisfying the method requirements. For large datasets, the higher computational requirement of this method might be prohibitive. This drawback can be partially suppressed with an alternative distance measure. We proposed to use Euclidean distance in combination with standard deviation normalization as a reliable alternative.

Keywords: outlier analysis, distance based method, global outlier, single cluster, Mahalanobis distance, biology

1 Introduction

Data mining reveals new, valuable and non-trivial information in large datasets [14]. It is a process of discovering interesting patterns and knowledge in the data that is not immediately apparent. Various data mining approaches help to specify the patterns in the data mining tasks. Examples include characterization and discrimination, mining of frequent patterns, associations and correlations, classification and regression, clustering analysis, and outlier analysis [10].

The outlier analysis has an important position among data mining approaches. Hawkins specified an intuitive definition of the term *outlier* as: 'Within a given dataset, the outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism' [11].

The other observations are usually called *inliers*, *normal data* or *normal observations*. Throughout the text, a predetermined battery of *features* characterizes an observation.

The outlier analysis is used in a wide variety of domains such as the financial industry, quality control, fault diagnosis, intrusion detection, web analytics, and medical diagnosis [2]. The most typical application of the outlier analysis is data cleaning. However, in many applications, outliers are more interesting than inliers. Fraud detection is a classic example, where attention focuses on the outliers, because these more likely represent cases of fraudulent behaviour [12].

The outlier analysis distinguishes three categories of outliers that require specific analytical approaches: *global outliers*, *contextual outliers* (known also as *conditional outliers*), and *collective outliers* [10]. A global outlier is an observation that deviates significantly from the rest of the dataset, whereas a contextual outlier deviates from inliers only with respect to a specific context. The term collective outlier is used for a subset of observations. A subset of observations forms a collective outlier if the subset as a whole deviates significantly from the entire dataset.

To identify outliers, the outlier detection methods create models of normal patterns in the data (so called *data model* or simply *model*), and then compute an *outlier score* of a given observation on the basis of the deviations from the normal patterns [2]. The outlier detection methods utilize *clustering models*, *distance-based models*, *density-based models*, *probabilistic* and *statistical models*, *classification models*, and *information-theoretic models* [10, 2].

The selection of the model and outlier score calculation is data-specific and relies on assumptions of information contained in the data. For example, classification models require datasets of labelled observations. Methods based on other models, e.g. statistical models or distance-based models, can be applied to both labelled as well as unlabelled datasets.

The correct choice of the method from the perspective of the data model determines results of the outlier analysis [2]. For example, application of a method based on a statistical model, which expects a uniform distribution of inliers, would be inappropriate for a dataset with the zipf distribution.

In biology, animal group behaviour studies generate specific datasets of observable variables pre-selected in the experimental design [5, 18]. Typically, such datasets are unlabelled and may contain numerical as well as categorical data. Given the complexity of animal behaviour, feature space of the observed variables will not be exhaustive on an individual level and determinants of group behaviour will exhibit subtle trends. From amongst the data mining approaches, the outlier analysis provides functionality to identify observations putatively generated by an alternative mechanism, which makes the analysis suitable for application in animal group behaviour research. In order to ensure a simple and reliable recognition of the outliers in such a dataset, we developed an outlier detection method. Our method detects global outliers using a distance-based model. Here, we introduce the method for numerical data.

2 Methods

2.1 Analysis of the Problem

A dataset considered for application with the proposed method contains inliers that form one multidimensional cluster, while the outliers span at a distance from the cluster centre. The outliers may or may not form small clusters. The total number of observations in the dataset range from tens to hundreds of observations. Further, distribution of inliers may significantly vary among various datasets. The data contains no prior knowledge about the outliers, and the information embodied in the outliers is the object of interest. These datasets may include both numerical and categorical data; however, the proposed method is intended for datasets composed of numerical data.

The first step in developing a new outlier detection method is identification of the outlier category. Following the above stated setup, the proposed method detects global outliers. The second step, selection of the model for inliers, delineates the direction of the development process. Herein, we use backward selection to select the proper model. Information-theoretic models are impropriate for the defined datasets, because of the expected type of features. Without prior knowledge about the outliers, the new detection method cannot be based on a classification model. As different datasets may have different distributions of inliers, usage of a probabilistic, density-based or a statistical model is inadvisable. Consequently, the method has to be based on one of the remaining model types; clustering or distance-based models.

Both clustering models and distance-based models represent appropriate choices for the new outlier detection method given the data. Between them, distance-based methods enable a higher granularity of analysis as compared to clustering methods. This property of distance-based methods provides a more refined ability to distinguish between weak and strong outliers in noisy data sets [2]. Hence, the presented method has been developed on a distance-based model.

2.2 Description of the Method

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n observations $\mathbf{x}$. The i-th observation $\mathbf{x}_i$, where $i \in I$ and $I = \{1, \ldots, n\}$, is a d-dimensional real vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ of features $x \in \mathcal{F}$, where x_{ik} is the k-th feature of the i-th observation, and $\mathcal{F}$ is a feature space. Let us expect that the majority of the observations, say m, belongs to inliers. The remaining p observations correspond to outliers. A subset of all outliers in the set X will be denoted as O.

The presented method belongs to the group of outlier detection methods based on a distance model. It assumes two attributes in the observations $\mathbf{x} \in X$:

(I) inliers form one predominant cluster and the outliers are at distances from the centre of the cluster,

(II) the number of inliers is much higher than the number of outliers ($m >> p$).

In order to design the separation method, the outlier score had to be properly formulized. For this purpose, an appropriate similarity measure had to be chosen. Similarity of two observations, say $\mathbf{x}_i$, $\mathbf{x}_j \in X$, was assessed using a distance measure. In order to ensure a comparable level of impact for all the features $x \in \mathcal{F}$, the observations should be compared with normalized data or the measure should be unitless and scale-invariant. In our solution, we used Mahalanobis distance [7, 4]. This measure is unitless and scale-invariant. For the observations $\mathbf{x}_i, \mathbf{x}_j$, the Mahalanobis distance is defined as

$$d\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)^\top}, \tag{1}$$

where $\mathbf{S}$ is a covariance matrix, and $\top$ symbolizes transposition.

Considering the properties of the datasets expressed via the assumptions (I) and (II), we proposed to formulate the outlier score J for the i-th observation as the sum of distances between the i-th observation and the others, i.e.

$$J_i = \sum_{\forall j \in I} d\left(\mathbf{x}_i, \mathbf{x}_j\right). \tag{2}$$

The distance-based methods usually take into account distances between an evaluated observation and its k nearest neighbours. Nevertheless, the outlier score (2) considers all n distances. An example demonstrates rationalization for the formulation of the score. In Fig. 1, the number of distances is identical regardless of whether an inlier (Fig. 1 a) or an outlier (Fig. 1 b) is evaluated; however, distributions of their values differ. For the outliers, longer distances appear more frequently than for inliers. This holds for an arbitrarily chosen inlier and outlier, since inliers form a single cluster and $m >> p$.

The specific properties of the dataset lead to the conclusion that the greater the number of nearest neighbours which are included in the analysis, the larger the difference between scores of inliers and outliers. Consequently, inclusion of all observations in the comparison results in higher sensitivity of the method. The associated increase in computational complexity of the method is irrelevant for the expected dataset sizes. For larger datasets, a GPU optimized variant of the method may be developed [3].

Observations evaluated using the outlier score (2) can be easily classified as outliers or inliers using a threshold value t. In our case, the unusual structure of the dataset X inspired the analytical expression of t. Indeed, values of the score for inliers are markedly smaller than for outliers. Considering this fact and the fact that $m >> p$, median of the score's values $\hat{J}$ adequately describes inliers. On the basis of the median and the smallest score values, the range of score values of inliers can be estimated. Thus, the threshold value can be expressed as

$$t = \varepsilon.[\hat{J} - \min_{\forall i \in I} J_i] + \hat{J}, \tag{3}$$

where the parameter ε is used as a tuning parameter. Each observation with a score equal to or greater than the threshold value t is expected to be an outlier.

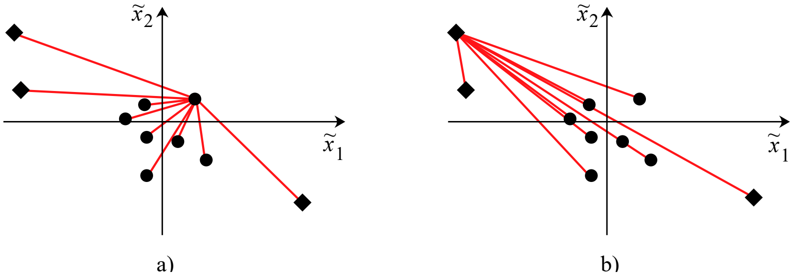


Fig. 1. Demonstration of the idea behind the outlier score by evaluation of: a) an inlier, and b) an outlier. In both figure panels, three outliers (diamonds) and seven inliers (circles) are plotted on a two-dimensional centred, rotated and standardized feature space $\tilde{x}_1, \tilde{x}_2 \in \tilde{\mathcal{F}}$. The distance between two observations is symbolized using a red line.

The presented method has one tuning parameter ε, and its setting significantly predetermines the output of the method. We proposed a visualization method in order to estimate the accurate value for ε. The visualization displays a continuous line connecting scores J for $\forall \mathbf{x} \in X$, where the scores are sorted in ascending order. The line is approximately exponential. The initial lag phase with gradual increase in J, includes inliers, and the subsequent exponential phase includes outliers. Using the graph, an expert can estimate the boundary between inliers and outliers and accordingly the threshold value determining ε (Fig. 2).

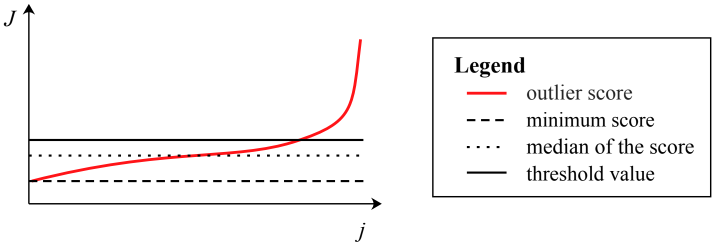


Fig. 2. Visualization of the score J as a function j where j are indexes of the observations sorted according to J in the ascending order. Threshold for outlier classification t is placed at a point with rapid change in score trend.

For datasets satisfying the assumptions, the right placement of the auxiliary line is straightforward. However, the more the dataset X deviates from the ideal, the deeper understanding of the data is necessary for the appropriate placement.

2.3 Algorithmic Expression of the Method

The proposed method can be realized as a function, here presented as a pseudocode (Algorithm 1). The function has two inputs and two outputs. The inputs are the set of observations X and the tuning parameter ε. The outputs are a set of all outliers O and a set of outliers indexes I_o in the original set X.

Algorithm 1 Extraction of Exceptional Observations

```
function EEO($X, \varepsilon$)
Input: Set of $n$ observations $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, constant $\varepsilon$ specifying the limit for outliers
Output: Set $O$ of all outliers and set of their indexes $I_o$ in $X$
    $J_i \leftarrow \sum_{\forall j \in I} d(\mathbf{x}_i, \mathbf{x}_j), \forall i \in I$    ▷ Evaluation of observations using the criterion
    $t \leftarrow \varepsilon.[\hat{J} - \min_{\forall i \in I} J_i] + \hat{J}$    ▷ Threshold value for exceptional observations
    $I_o \leftarrow \{i : i \in I \text{ where } J_i \geq t\}$    ▷ Indexes of exceptional observations
    $O \leftarrow \{\mathbf{x}_i : i \in I_o\}$    ▷ Set of exceptional observations
    return $O, I_o$
end function
```

2.4 Experimental Evaluation of the Method

We used 44 previously published datasets for evaluation of the proposed method [8]. The datasets originated from areas such as biology, medicine, criminology or astronautics. They contained three types of features: R - real numbers, I - integers, and N - nominal values. The datasets consisted of labelled samples with two classes. All datasets were imbalanced with an imbalance ratio $IR = m/p$, where $IR \in [1.82, 129.44]$. We expected that the minority class represented outliers, while the majority class consisted of inliers.

We adapted three performance measures used in binary classification to evaluate results obtained from our method. Namely, we considered *sensitivity* (Se), *specificity* (Sp), and their geometric mean (G) [8, 15]. For the outlier analysis, they can be expressed as

$$Se = \frac{|TO|}{|TO| + |FI|}, \qquad Sp = \frac{|TI|}{|TI| + |FO|}, \qquad G = \sqrt{Se \cdot Sp}, \tag{4}$$

where $|TO|$ is the number of correctly recognized outliers (true outliers), $|FO|$ is the number of inliers labelled as outliers (false outliers), $|TI|$ is the number of correctly recognized inliers (true inliers), $|FI|$ is the number of outliers labelled as inliers (false inliers).

We evaluated our method (extraction of exceptional observations, EEO) for two values of ε. Within the first experiment, we estimated the value of ε from the graph (as per Fig. 2). This value was denoted as $\hat{\varepsilon}$. In the second experiment, we searched for an optimal setting (ε^*) using genetic algorithms [16]. The genetic algorithms used the objective function as $\max G(\varepsilon)$. We applied the MATLAB function `ga`, with no constraints and default settings [1].

3 Results

Due to the presence of nominal variables, EEO could not be applied on datasets 'abalone19' and 'abalone9-18'. Further, it was unsuccessful on datasets 'ecoli-0_vs_1' and 'segment0', in which some features had constant values for all observations. The obtained results are summarized in Table 1. In general, sensitivity of EEO with manual estimation of $\hat{\varepsilon}$ was lower than for ε^*, established from labelled data with genetic algorithms in lieu of higher specificity. This was accompanied by higher, and thus more conservative, values of $\hat{\varepsilon}$.

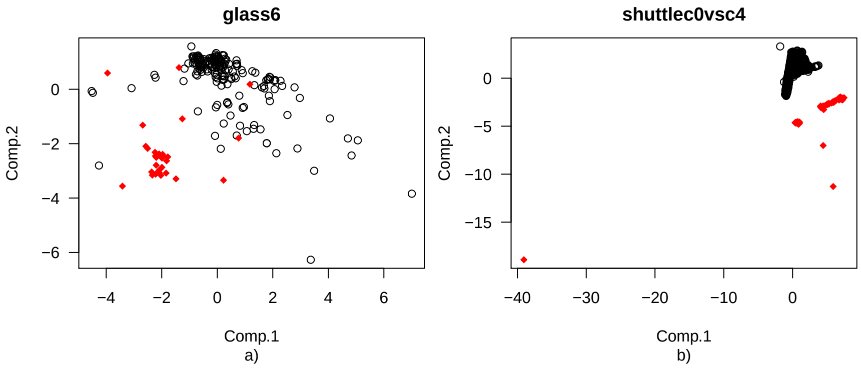
To estimate the performance of EEO amongst existing outlier detection methods, we compared our method with Chi et al.'s method with 3 and 5 labels (Chi-3 and Chi-5) [6], Ishibuchi et al.'s method (Ish05) [13], E-Algorithm (E-Alg) [19], Fernández et al.'s method (HFRBCS) [8], and C4.5 decision tree (C4.5) [17]. We adopted the evaluation results published in [8]. The evaluation results using G are summarized for all expected methods, including the EEO with optimal and manual setting of ε, in Table 2.

4 Discussion

The proposed EEO method was tested on 44 datasets previously used for algorithm testing [8]. The datasets differed in the imbalance ratio, in the number of features and their type (Table 1). However, from the viewpoint of EEO testing, many of these datasets did not meet the assumptions that the inliers form one multidimensional cluster (I) and the number of inliers is much higher than the number of inliers (II). This is apparent when displaying the first two principal components of observation scores with their class labels [9]. The dataset 'shuttle-c0-vs-c4' fully met the assumptions (Fig. 3 b), and EEO was successful in outlier detection ($G \approx 98$). The datasets 'ecoli-0-1-3-7_vs_2-6', 'shuttle-c2-vs-c4', and 'Wisconsin' similarly showed nearly ideal class assignment with respect to inliers (data not shown). On these datasets, EEO exhibited excellent results according to all three measures (4) both for $\hat{\varepsilon}$ and ε^*. The performance of EEO is fully comparable to all evaluated methods. In fact, EEO provides considerably better separation on 'ecoli-0-1-3-7_vs_2-6' dataset than any other considered method (Table 2).

Good results were obtained also on other datasets, e.g. on 'glass6' (Fig. 3 a), 'new-thyroid1', or 'yeast-2_vs_8'. Here, a majority of the inliers were concentrated near the cluster center; however, many inliers (their number was similar to the total number of outliers) were interspersed with the outliers. In such cases, estimation of ε became vague and perfect separation was not possible. Thus, the good EEO results on these datasets were coincidental and the presented method was not suited for them.

While the threshold values t for outlier detection can be directly set from the sorted distance visualization, estimating ε will represent a good practice in data reporting. The ε value defines the position of the outliers relative to the median, providing a data-independent approximation on outlier distribution comparable between studies.



a) b)

Fig. 3. Example of a) inappropriate and b) ideal datasets. The outliers (red diamonds) and inliers (black circles) are plotted in a two-dimensional centred, rotated and standardized feature space using the first two principal components.

The presented method was based on the Mahalanobis distance (1). While the distance was efficient for the proposed problem, we found the method to be computationally extravagant. Thus, we suggest an alternative approach based on the Euclidean distance for application where computational intensity would be of concern. The Euclidean distance in combination with standard deviation normalization [14] might provide equally good results while its time-complexity would be considerably lower.

5 Conclusion

The outlier analysis has the potential to mine valuable information from a complex dataset, but its sensitivity and specificity is dependent on both suitability of the method and the model, to the data. We designed EEO for the specifics of animal group behaviour observations, where the outliers could reveal alternative mechanisms determining group behaviour. Our testing on varied imbalanced sets demonstrated that the utility of the method is wider. The EEO was able to correctly classify outliers in datasets from engineering, microbiology or medicine. We therefore conclude that global outliers may be detected with EEO based on the threshold estimated from sums of pairwise Mahalanobis distances in datasets across fields that form one predominant multidimensional cluster with outliers distanced from it.

Acknowledgement: The work was supported by the University of Pardubice (PŠ) and the Czech Science Foundation grant number 17-20286S (NM).

References

1. MATLAB: Global optimization toolbox (R2016a) (2016), https://www.mathworks.com/help/gads/index.html
2. Aggarwal, C.C.: Outlier Analysis. Springer New York (2013)
3. Angiulli, F., Basta, S., Lodi, S., Sartori, C.: GPU strategies for distance-based outlier detection. IEEE Transactions on Parallel and Distributed Systems 27(11), 3256–3268 (Nov 2016)
4. Brereton, R.G.: The mahalanobis distance and its relationship to principal component scores. Journal of Chemometrics 29(3), 143–145 (2015)
5. Broom, D.M., Fraser, A.F.: Domestic animal behaviour and welfare. Cabi, 4th edn. (2015)
6. Chi, Z., Yan, H., Pham, T.: Fuzzy algorithms: with applications to image processing and pattern recognition, vol. 10. World Scientific (1996)
7. Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer-Verlag Berlin Heidelberg, 3rd edn. (2014)
8. Fernndez, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. International Journal of Approximate Reasoning 50(3), 561–577 (2009)
9. Gower, J., Lubbe, S., Roux, N.: Understanding Biplots. John Wiley & Sons, Inc. (2010)
10. Han, J., Kamber, M., Pei, J.: Data Mining. Morgan Kaufmann, 3rd edn. (2012)
11. Hawkins, D.M.: Identification of Outliers. Springer Netherlands (1980)
12. Hawkins, S., He, H., Williams, G., Baxter, R.: Outlier detection using replicator neural networks. In: Kambayashi, Y., Winiwarter, W., Arikawa, M. (eds.) Data Warehousing and Knowledge Discovery: 4th International Conference, DaWaK 2002 Aix-en-Provence, France, September 4–6, 2002 Proceedings. pp. 170–180. Springer Berlin Heidelberg, Berlin, Heidelberg (2002)
13. Ishibuchi, H., Yamamoto, T.: Rule weight specification in fuzzy rule-based classification systems. IEEE transactions on fuzzy systems 13(4), 428–435 (2005)
14. Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Inc., 2nd edn. (2011)
15. Kohl, M.: Performance measures in binary classification. International Journal of Statistics in Medical Research 1(1), 79–81 (2012)
16. Reeves, C.R., Rowe, J.E.: Genetic Algorithms: Principles and Perspectives: A Guide to GA Theory. Kluwer Academic Publishers, Norwell, MA, USA (2002)
17. Salzberg, S.L.: C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. Machine Learning 16(3), 235–240 (1994)
18. Ward, A., Webster, M.: Sociality: The Behaviour of Group-Living Animals. Springer International Publishing (2016)
19. Xu, L., y. Chow, M., Taylor, L.S.: Using the data mining based fuzzy classification algorithm for power distribution fault cause identification with imbalanced data. In: 2006 IEEE PES Power Systems Conference and Exposition. pp. 1228–1233 (Oct 2006)

Table 1. Evaluation of EEO on test datasets using sensitivity Se, specificity Sp, and their geometric mean G. In first four columns, basic information about datasets is listed. It includes name, feature type (R - real numbers, I - integers, and N - nominal values), number of observations n, and imbalance ratio IR. The remaining columns consist of evaluation results for estimated and suboptimal setting of ε, respectively.

Information about datasets				EEO with $\hat{\varepsilon}$				EEO with ε^*			
Name	R/I/N	n	IR	ε	Se	Sp	G	ε	Se	Sp	G
abalone19	7/0/1	4174	129.44	-	-	-	-	-	-	-	-
abalone9-18	7/0/1	472	16.4	-	-	-	-	-	-	-	-
ecoli-0_vs_1	7/0/0	220	1.86	-	-	-	-	-	-	-	-
ecoli-0-1-3-7_vs_2-6	7/0/0	281	39.14	0.60	71.43%	86.13%	78.44%	2.47	71.43%	98.91%	84.05%
ecoli1	7/0/0	336	3.36	0.70	20.78%	83.78%	41.72%	0.04	59.74%	55.98%	57.83%
ecoli2	7/0/0	336	5.46	0.90	9.62%	86.97%	28.92%	0.26	57.69%	64.79%	61.14%
ecoli3	7/0/0	336	8.6	0.90	8.57%	87.04%	27.31%	0.06	65.71%	55.15%	60.20%
ecoli4	7/0/0	336	15.8	0.90	65.00%	90.82%	76.83%	0.75	70.00%	87.34%	78.19%
glass0	9/0/0	214	2.06	1.30	10.00%	68.75%	26.22%	-0.30	51.43%	25.69%	36.35%
glass-0-1-2-3_vs_4-5-6	9/0/0	214	3.2	1.30	52.94%	84.66%	66.95%	0.88	86.27%	83.44%	84.84%
glass-0-1-6_vs_2	9/0/0	192	10.29	1.60	17.65%	77.71%	37.03%	-0.21	88.24%	37.71%	57.69%
glass-0-1-6_vs_5	9/0/0	184	19.44	1.40	22.22%	78.86%	41.86%	0.85	55.56%	69.71%	62.23%
glass1	9/0/0	214	1.82	1.30	19.74%	73.19%	38.01%	-0.22	56.58%	34.78%	44.36%
glass2	9/0/0	214	11.59	1.00	11.76%	85.28%	31.67%	0.12	41.18%	55.84%	47.95%
glass4	9/0/0	214	15.47	1.30	46.15%	77.11%	59.66%	0.59	92.31%	66.67%	78.45%
glass5	9/0/0	214	22.78	1.30	22.22%	75.61%	40.99%	0.88	55.56%	67.80%	61.38%
glass6	9/0/0	214	6.38	1.30	65.52%	82.16%	73.37%	0.87	96.55%	76.76%	86.09%
haberman	0/3/0	306	2.78	0.60	14.81%	95.11%	37.54%	0.13	46.91%	66.22%	55.74%
iris0	4/0/0	150	2	0.80	10.00%	81.00%	28.46%	-0.22	86.00%	30.00%	50.79%
new-thyroid1	4/1/0	215	5.14	0.70	65.71%	83.89%	74.25%	0.31	80.00%	74.44%	77.17%
new-thyroid2	4/1/0	215	5.14	0.70	65.71%	83.89%	74.25%	0.23	82.86%	71.11%	76.76%
page-blocks0	4/6/0	5472	8.79	0.70	85.87%	83.88%	84.87%	0.69	86.05%	83.68%	84.85%
page-blocks-1-3_vs_4	4/6/0	472	15.86	0.70	53.57%	78.60%	64.89%	0.46	71.43%	71.40%	71.41%
pima	8/0/0	768	1.87	0.90	20.15%	88.40%	42.20%	0.10	61.57%	67.20%	64.32%
segment0	19/0/0	2308	6.02	-	-	-	-	-	-	-	-
shuttle-c0-vs-c4	0/9/0	1829	13.87	0.70	100.00%	95.96%	97.96%	0.77	100.00%	97.01%	98.49%
shuttle-c2-vs-c4	0/9/0	129	20.5	0.50	100.00%	88.62%	94.14%	0.97	100.00%	96.75%	98.36%
vehicle0	0/18/0	846	3.25	0.50	16.08%	90.42%	38.13%	0.05	53.27%	60.59%	56.81%
vehicle1	0/18/0	846	2.9	0.40	9.68%	82.83%	28.31%	-0.02	48.85%	46.10%	47.46%
vehicle2	0/18/0	846	2.88	0.40	18.35%	85.83%	39.68%	0.14	31.65%	65.61%	45.57%
vehicle3	0/18/0	846	2.99	0.40	10.85%	83.28%	30.06%	-0.12	70.75%	37.54%	51.54%
vowel0	10/3/0	988	9.98	0.40	48.89%	86.64%	65.08%	0.20	68.89%	72.72%	70.78%
wisconsin	0/9/0	683	1.86	0.50	97.07%	88.51%	92.69%	1.15	93.72%	94.37%	94.05%
yeast-0-5-6-7-9_vs_4	8/0/0	528	9.35	0.50	37.25%	81.34%	55.05%	0.30	54.90%	72.54%	63.11%
yeast1	8/0/0	1484	2.46	0.50	18.41%	78.58%	38.04%	-0.03	48.48%	45.31%	46.87%
yeast-1_vs_7	7/0/0	459	14.3	0.70	40.00%	85.31%	58.42%	0.24	53.33%	66.67%	59.63%
yeast-1-2-8-9_vs_7	8/0/0	947	30.57	0.70	40.00%	83.32%	57.73%	0.39	53.33%	72.30%	62.10%
yeast-1-4-5-8_vs_7	8/0/0	693	22.1	0.70	10.00%	82.50%	28.72%	-0.13	73.33%	40.57%	54.55%
yeast-2_vs_4	8/0/0	514	9.08	0.60	50.98%	84.45%	65.61%	0.18	84.31%	69.76%	76.69%
yeast-2_vs_8	8/0/0	482	23.1	0.60	70.00%	81.82%	75.68%	1.20	65.00%	93.94%	78.14%
yeast3	8/0/0	1484	8.1	0.50	25.15%	80.02%	44.86%	-0.01	69.33%	51.40%	59.69%
yeast4	8/0/0	1484	28.1	0.50	45.10%	80.32%	60.19%	0.16	78.43%	62.11%	69.79%
yeast5	8/0/0	1484	32.73	0.70	34.09%	86.74%	54.38%	0.06	100.00%	55.69%	74.63%
yeast6	8/0/0	1484	41.4	0.70	22.86%	86.34%	44.42%	-0.04	91.43%	47.83%	66.13%

Table 2. Comparison of EEO with other approaches for outlier detection. The geometric mean G of sensitivity and specificity was used as an overall comparison value. Results obtained by EEO are in bold on relevant datasets that meet designed criteria (inliers form a predominant cluster with outliers spanned from it and the number of inliers is greater than the number of outliers).

Dataset	Chi-3	Chi-5	Ish05	E-Alg	HFRBCS	C4.5	EEO	
							$\hat{\varepsilon}$	ε^*
abalone19	62.69%	66.71%	66.09%	0.00%	70.19%	15.58%	-	-
abalone9-18	63.93%	66.47%	65.78%	32.29%	67.56%	53.19%	-	-
ecoli-0_vs_1	92.27%	95.56%	96.70%	95.25%	93.63%	67.95%	-	-
ecoli-0-1-3-7_vs_2-6	71.04%	49.57%	71.31%	73.65%	71.48%	71.21%	**78.44**%	**84.05**%
ecoli1	85.28%	86.05%	85.71%	77.81%	84.18%	76.10%	41.72%	57.83%
ecoli2	88.01%	87.64%	87.00%	70.35%	87.62%	91.60%	28.92%	61.14%
ecoli3	87.58%	91.61%	85.39%	78.54%	90.81%	88.77%	27.31%	60.20%
ecoli4	91.27%	92.11%	86.92%	92.43%	93.02%	81.28%	76.83%	78.19%
glass0	64.06%	63.69%	69.39%	0.00%	76.57%	78.14%	26.22%	36.35%
glass-0-1-2-3_vs_4-5-6	85.83%	85.94%	88.56%	82.09%	88.37%	90.13%	66.95%	84.84%
glass-0-1-6_vs_2	40.84%	56.17%	41.18%	0.00%	58.37%	48.91%	37.03%	57.69%
glass-0-1-6_vs_5	71.48%	75.59%	88.77%	65.14%	77.96%	72.08%	41.86%	62.23%
glass1	64.90%	64.91%	59.29%	0.00%	73.66%	75.11%	38.01%	44.36%
glass2	47.67%	49.24%	43.55%	9.87%	54.84%	33.86%	31.67%	47.95%
glass4	84.96%	81.75%	78.27%	83.38%	70.39%	83.71%	59.66%	78.45%
glass5	81.56%	64.33%	89.96%	50.61%	68.73%	86.70%	40.99%	61.38%
glass6	83.87%	78.13%	86.27%	90.23%	86.95%	83.00%	73.37%	86.09%
haberman	58.91%	60.40%	62.65%	4.94%	57.08%	61.32%	37.54%	55.74%
iris0	100.00%	98.97%	100.00%	100.00%	100.00%	98.97%	28.46%	50.79%
new-thyroid1	87.44%	95.38%	89.02%	88.52%	95.58%	97.98%	74.25%	77.17%
new-thyroid2	89.81%	96.34%	94.21%	88.57%	99.72%	96.51%	74.25%	76.76%
page-blocks0	79.91%	87.25%	32.16%	64.51%	91.40%	94.84%	84.87%	84.85%
page-blocks-1-3_vs_4	91.92%	92.93%	94.53%	94.12%	98.64%	99.55%	64.89%	71.41%
pima	66.80%	66.78%	71.10%	55.01%	68.72%	71.26%	42.20%	64.32%
segment0	94.99%	95.88%	42.47%	95.33%	97.51%	99.26%	-	-
shuttle-c0-vs-c4	99.12%	98.71%	99.16%	98.40%	99.12%	99.97%	**97.96**%	**98.49**%
shuttle-c2-vs-c4	89.99%	78.34%	99.17%	100.00%	97.49%	99.15%	**94.14**%	**98.36**%
vehicle0	86.41%	84.93%	75.94%	39.07%	88.92%	91.10%	38.13%	56.81%
vehicle1	70.92%	71.88%	64.89%	3.09%	71.76%	69.28%	28.31%	47.46%
vehicle2	85.54%	87.19%	67.82%	43.83%	90.61%	94.85%	39.68%	45.57%
vehicle3	69.22%	63.13%	63.12%	0.00%	66.80%	74.34%	30.06%	51.54%
vowel0	98.37%	97.87%	89.03%	89.63%	98.82%	94.74%	65.08%	70.78%
wisconsin	88.91%	43.58%	95.78%	96.01%	88.24%	95.44%	**92.69**%	**94.05**%
yeast-0-5-6-7-9_vs_4	78.91%	75.99%	79.49%	59.99%	73.18%	74.88%	55.05%	63.11%
yeast1	67.69%	69.66%	51.41%	0.00%	71.71%	70.86%	38.04%	46.87%
yeast-1_vs_7	80.05%	63.02%	53.15%	27.55%	70.74%	67.73%	58.42%	59.63%
yeast-1-2-8-9_vs_7	76.12%	69.26%	48.55%	50.00%	69.37%	64.13%	57.73%	62.10%
yeast-1-4-5-8_vs_7	62.40%	58.76%	40.80%	0.00%	62.49%	41.19%	28.72%	54.55%
yeast-2_vs_4	86.80%	86.39%	70.85%	80.92%	89.32%	85.09%	65.61%	76.69%
yeast-2_vs_8	72.75%	78.76%	72.83%	72.83%	72.47%	78.23%	75.68%	78.14%
yeast3	90.13%	89.33%	77.06%	81.99%	90.41%	88.50%	44.86%	59.69%
yeast4	82.99%	83.07%	71.36%	32.16%	82.64%	65.00%	60.19%	69.79%
yeast5	93.41%	93.64%	94.94%	88.17%	94.20%	92.04%	54.38%	74.63%
yeast6	87.50%	87.73%	88.42%	51.72%	84.92%	80.38%	44.42%	66.13%