ECOD

Баранов В.М., Кисляков Г.И., Орлов Г.Е., Румшевич А.Д., Родин А.М.

Введение

ECOD - метод, использующий эмпирические функции кумулятивного распределения для обнаружения выбросов.

Преимущества метода:

- Не имеет гиперпараметров
- Обладает временной сложностью, которая линейно масштабируется в зависимости от размера набора данных и количества измерений
- Эффективность и масштабируемость
- Интерпертируемость

Принцип работы ECOD

$X_1, X_2, \ldots, X_n \in \mathbb{R}^d; \quad X \in \mathbb{R}^{n \times d}$

Далее через $y^{(j)}$ ссылаемся на, соответственно, j-ую компоненту вектора y

Введём CDF - Функцию распределения СВ:

$F : \mathbb{R}^d \to [0, 1]; \; F(x) = \mathbb{P}(X^{(1)} \leq x^{(1)}, \ldots, X^{(d)} \leq x^{(d)})$

$$F(x) = \prod_j F^{(j)}(x^{(j)}); \quad F^{(j)}(z) = \mathbb{P}(X^{(j)} \leq z)$$

Оценивается в алгоритме с помощью ECDF(Эмпирической CDF):

$$\widehat{F}^{(j)}(z) = \frac{1}{n} \sum_i \mathbb{I}(X_i^{(j)} \leq z)$$

left tail ECDF: $\widehat{F}^{(j)}_{\text{left}}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i^{(j)} \leq z\}$ for $z \in \mathbb{R}$,

right tail ECDF: $\widehat{F}^{(j)}_{\text{right}}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i^{(j)} \geq z\}$ for $z \in \mathbb{R}$.

$$\gamma_j = \frac{\frac{1}{n}\sum_{i=1}^{n}\big(X_i^{(j)} - \overline{X^{(j)}}\big)^3}{\Big[\frac{1}{n-1}\sum_{i=1}^{n}\big(X_i^{(j)} - \overline{X^{(j)}}\big)^2\Big]^{3/2}},$$

$$O_{\text{left-only}}(X_i) := -\log \widehat{F}_{\text{left}}(X_i) = -\sum_{j=1}^{d} \log(\widehat{F}_{\text{left}}^{(j)}(X_i^{(j)})),$$

(4)

$$O_{\text{right-only}}(X_i) := -\log \widehat{F}_{\text{right}}(X_i) = -\sum_{j=1}^{d} \log(\widehat{F}_{\text{right}}^{(j)}(X_i^{(j)})).$$

$$O_{\text{auto}}(X_i) = -\sum_{j=1}^{d} [\mathbb{1}\{\gamma_j < 0\} \log(\widehat{F}_{\text{left}}^{(j)}(X_i^{(j)})) + \mathbb{1}\{\gamma_j \geq 0\} \log(\widehat{F}_{\text{right}}^{(j)}(X_i^{(j)})).$$

$$O_i = \max\{O_{\text{left-only}}(X_i), O_{\text{right-only}}(X_i), O_{\text{auto}}(X_i)\}.$$

Свойства ECOD

1) Интерпретируемый детектор выбросов

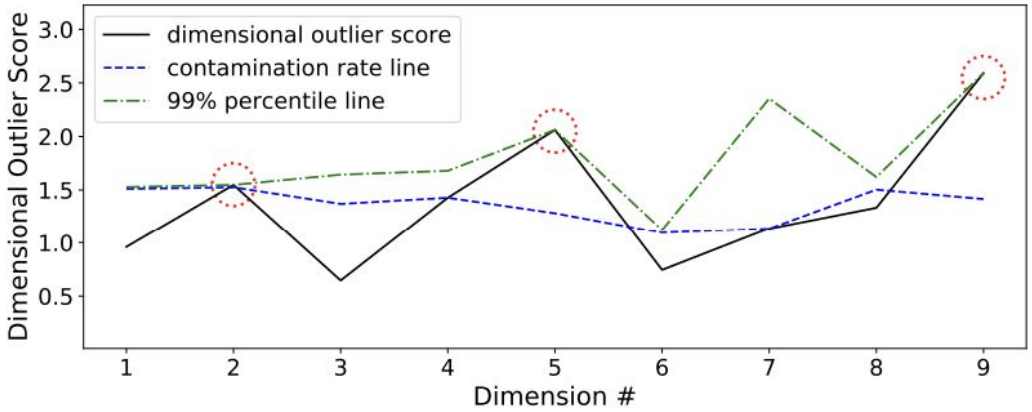


Fig. 2: *Dimensional Outlier Graph* for sample 70 (classified as an outlier) of BreastW dataset; dimension 2, 5, 9 are identified with most contribution to its outlyingness.

Свойства ECOD

2) Масштабируемость

3) Многопоточность

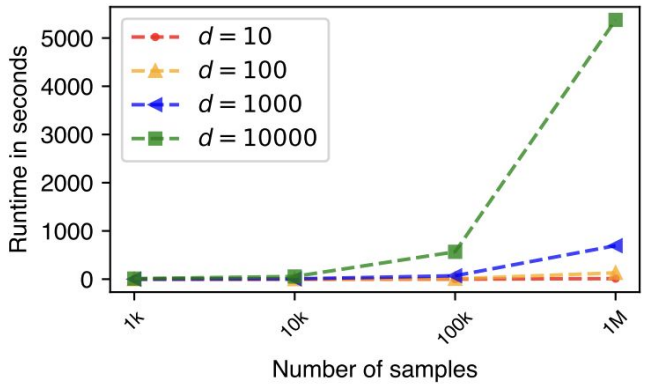


Fig. 5: Runtime of ECOD on the synthetic datasets with a varying number of samples and dimensions. It is efficient and scalable to handle a million samples with 10,000 dimensions in less than 2 hours.

TABLE 6: ECOD's runtime (in seconds) on a moderate laptop under different sample size (n) and dimensions (d). It scales well to handle high-dimensional, large datasets.

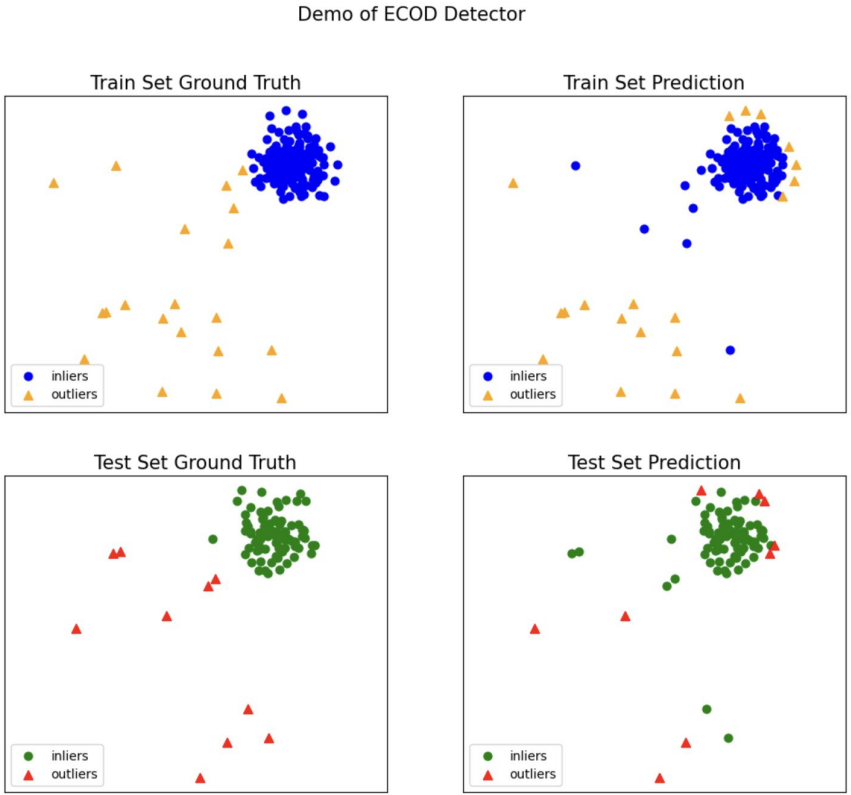
	d=10	d=100	d=1,000	d=10,000
n=1,000	0.068	0.173	1.163	11.460
n=10,000	0.171	0.468	5.244	55.190
n=100,000	0.640	7.185	70.541	567.105
n=1,000,000	11.403	130.974	694.405	5376.593

Реализация метода в Python

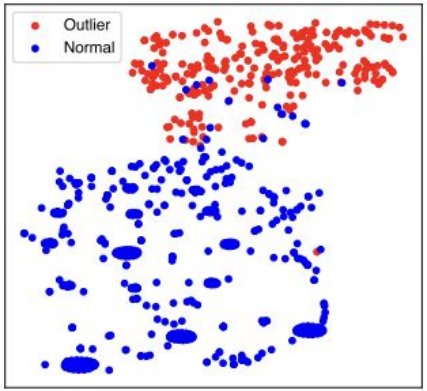
```
sys.path.append(
    os.path.abspath(os.path.join(os.path.dirname("__file__"), '..')))

from pyod.models.ecod import ECOD
from pyod.utils.data import generate_data
from pyod.utils.data import evaluate_print
from pyod.utils.example import visualize

if __name__ == "__main__":
    contamination = 0.1  # percentage of outliers
    n_train = 200  # number of training points
    n_test = 100  # number of testing points

    # Generate sample data
    X_train, X_test, y_train, y_test = \
        generate_data(n_train=n_train,
                      n_test=n_test,
                      n_features=2,
                      contamination=contamination,
                      random_state=42)

    # train ECOD detector
    clf_name = 'ECOD'
    clf = ECOD()

    # you could try parallel version as well.
    # clf = ECOD(n_jobs=2)
    clf.fit(X_train)

    # get the prediction labels and outlier scores of the training data
    y_train_pred = clf.labels_  # binary labels (0: inliers, 1: outliers)
    y_train_scores = clf.decision_scores_  # raw outlier scores

    # get the prediction on the test data
    y_test_pred = clf.predict(X_test)  # outlier labels (0 or 1)
    y_test_scores = clf.decision_function(X_test)  # outlier scores

    # evaluate and print the results
    print("\nOn Training Data:")
    evaluate_print(clf_name, y_train, y_train_scores)
    print("\nOn Test Data:")
    evaluate_print(clf_name, y_test, y_test_scores)

    # visualize the results
    visualize(clf_name, X_train, y_train, X_test, y_test, y_train_pred,
              y_test_pred, show_figure=True, save_figure=False)
```

```
On Training Data:
ECOD ROC:0.9562, precision @ rank n:0.65

On Test Data:
ECOD ROC:0.935, precision @ rank n:0.4
```

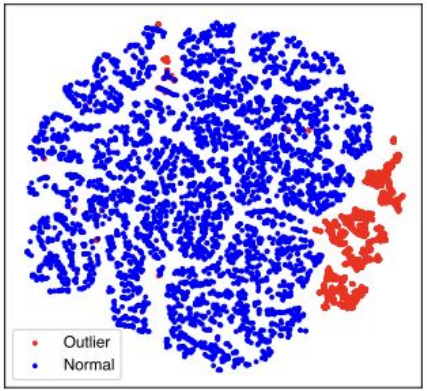
Реализация метода в Python



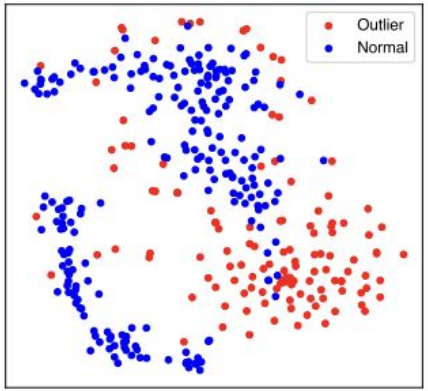
Как ECOD справляется с разными наборами данных



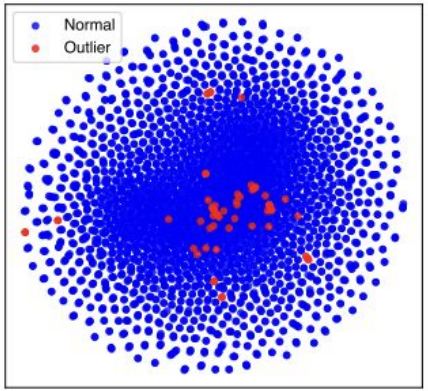
(a) *Breastw (mat)*



(b) *Shuttle (mat)*



(c) *Ionosphere (mat)*



(d) *Speech (mat)*

Сравнение с другими методами (средняя точность)

Data	ABOD	CBLOF	HBOS	IForest	KNN	LODA	LOF	LSCP	OCSVM	PCA	SUOD	ECOD
Arrhythmia (mat)	0.359 (12)	0.399 (9)	**0.493 (3)**	**0.506 (1)**	0.397 (10)	0.436 (5)	0.374 (11)	0.411 (6)	0.405 (7)	0.402 (8)	0.495 (2)	0.473 (4)
Breastw (mat)	0.295 (12)	0.914 (8)	0.953 (5)	0.972 (3)	0.927 (7)	0.978 (2)	0.322 (11)	0.826 (9)	0.934 (6)	0.96 (4)	0.791 (10)	**0.988 (1)**
Cardio (mat)	0.194 (11)	0.414 (8)	0.46 (6)	0.576 (3)	0.345 (10)	0.424 (7)	0.163 (12)	0.396 (9)	0.532 (4)	**0.611 (1)**	0.462 (5)	0.579 (2)
Ionosphere (mat)	0.914 (2)	0.871 (3)	0.366 (12)	0.789 (8)	**0.924 (1)**	0.716 (11)	0.821 (6)	0.818 (7)	0.823 (5)	0.736 (9)	0.824 (4)	0.719 (10)
Lympho (mat)	0.517 (11)	0.808 (9)	**0.925 (2)**	**0.952 (1)**	0.82 (7)	0.417 (12)	0.825 (6)	0.861 (4)	0.814 (8)	0.852 (5)	0.799 (10)	0.894 (3)
Mammography (mat)	0.023 (12)	0.137 (9)	0.122 (10)	0.233 (3)	0.17 (6)	0.281 (2)	0.118 (11)	0.169 (7)	0.188 (5)	0.199 (4)	0.164 (8)	**0.429 (1)**
Optdigits (mat)	0.028 (8)	0.062 (2)	**0.196 (1)**	0.055 (3)	0.022 (12)	0.024 (11)	0.029 (7)	0.043 (6)	0.028 (8)	0.028 (8)	0.046 (5)	0.053 (4)
Pima (mat)	0.511 (4)	0.477 (7)	**0.569 (1)**	0.503 (5)	**0.515 (3)**	0.408 (12)	0.43 (11)	0.47 (9)	0.461 (10)	0.478 (6)	0.473 (8)	0.541 (2)
Satellite (mat)	0.397 (11)	**0.69 (1)**	**0.687 (2)**	0.654 (3)	0.543 (9)	0.581 (8)	0.39 (12)	0.51 (10)	0.653 (4)	0.603 (6)	0.608 (5)	0.585 (7)
Satimage-2 (mat)	0.187 (11)	**0.978 (1)**	0.758 (7)	0.929 (3)	0.419 (9)	0.87 (5)	0.027 (12)	0.333 (10)	0.975 (2)	0.874 (4)	0.623 (8)	0.86 (6)
Shuttle (mat)	0.171 (11)	0.195 (10)	0.98 (3)	**0.986 (1)**	0.204 (9)	0.379 (8)	0.142 (12)	0.715 (7)	0.902 (6)	0.926 (4)	0.913 (5)	**0.981 (2)**
Speech (mat)	**0.04 (1)**	0.022 (6)	0.027 (3)	0.018 (11)	0.022 (6)	0.017 (12)	0.024 (5)	0.027 (3)	0.021 (9)	0.022 (6)	0.029 (2)	0.02 (10)
Wbc (mat)	0.355 (12)	0.5 (11)	0.663 (2)	0.59 (4)	0.529 (9)	0.564 (5)	0.558 (6)	0.554 (7)	0.514 (10)	0.534 (8)	0.602 (3)	**0.783 (1)**
Wine (mat)	0.084 (11)	0.06 (12)	0.405 (2)	0.279 (6)	0.095 (10)	0.278 (7)	0.361 (4)	0.299 (5)	0.141 (9)	0.254 (8)	0.364 (3)	**0.608 (1)**
Arrhythmia (arff)	0.699 (11)	0.712 (8)	0.75 (3)	0.746 (4)	0.712 (8)	0.697 (12)	0.704 (10)	0.718 (5)	0.716 (6)	0.714 (7)	**0.751 (2)**	**0.752 (1)**
Cardiotocography (arff)	0.247 (12)	0.363 (9)	0.366 (8)	0.434 (2)	0.311 (10)	0.432 (3)	0.258 (11)	0.374 (7)	0.419 (4)	**0.478 (1)**	0.398 (5)	0.378 (6)
HeartDisease (arff)	0.534 (4)	0.521 (9)	**0.625 (2)**	0.534 (4)	0.538 (3)	0.445 (12)	0.478 (11)	0.525 (8)	0.513 (10)	0.53 (6)	0.528 (7)	**0.64 (1)**
Hepatitis (arff)	0.33 (12)	0.374 (11)	0.473 (6)	0.442 (8)	0.475 (5)	0.396 (10)	0.499 (4)	0.472 (7)	0.427 (9)	0.543 (3)	0.557 (2)	**0.585 (1)**
InternetAds (arff)	0.276 (10)	0.315 (8)	**0.535 (1)**	0.49 (3)	0.281 (9)	0.242 (12)	0.262 (11)	0.387 (5)	0.316 (7)	0.32 (6)	0.475 (4)	0.51 (2)
Ionosphere (arff)	**0.915 (1)**	0.862 (3)	0.364 (12)	0.777 (8)	**0.915 (1)**	0.763 (9)	0.811 (7)	0.819 (6)	0.822 (5)	0.723 (10)	0.835 (4)	0.703 (11)
KDDCup99 (arff)	0.018 (12)	**0.198 (5)**	**0.278 (1)**	**0.273 (2)**	0.046 (10)	0.135 (7)	0.028 (11)	0.127 (8)	0.125 (9)	**0.199 (4)**	0.157 (6)	**0.239 (3)**
Lymphography (arff)	0.801 (11)	0.925 (7)	0.978 (3)	**0.992 (1)**	0.942 (6)	0.491 (12)	0.925 (7)	0.922 (9)	0.837 (10)	0.953 (4)	0.944 (5)	**0.982 (2)**
Pima (arff)	0.506 (5)	0.484 (7)	0.528 (2)	0.515 (4)	**0.525 (3)**	0.42 (12)	0.458 (11)	0.48 (9)	0.478 (10)	0.484 (7)	0.487 (6)	**0.53 (1)**
Shuttle (arff)	0.263 (5)	0.358 (3)	0.08 (12)	0.13 (10)	0.36 (2)	0.132 (9)	**0.395 (1)**	0.184 (7)	0.304 (4)	0.232 (6)	0.171 (8)	0.085 (11)
SpamBase (arff)	0.38 (10)	0.414 (8)	0.525 (2)	0.492 (3)	0.422 (6)	0.375 (11)	0.349 (12)	0.45 (4)	0.412 (9)	0.42 (7)	0.431 (5)	**0.567 (1)**
Stamps (arff)	0.247 (10)	0.241 (11)	0.398 (5)	0.394 (6)	0.341 (9)	0.404 (4)	0.229 (12)	0.425 (2)	0.346 (8)	0.411 (3)	0.392 (7)	**0.464 (1)**
Waveform (arff)	0.066 (8)	0.129 (2)	0.055 (10)	0.056 (9)	**0.134 (1)**	0.052 (12)	0.108 (4)	0.114 (3)	0.07 (6)	0.054 (11)	0.07 (6)	**0.079 (5)**
WBC (arff)	0.542 (11)	0.556 (9)	0.694 (6)	**0.863 (1)**	0.581 (8)	0.724 (4)	0.33 (12)	0.543 (10)	0.702 (5)	0.638 (7)	0.728 (3)	0.838 (2)
WDBC (arff)	0.43 (12)	0.667 (9)	0.795 (2)	0.718 (6)	0.654 (10)	0.794 (3)	0.704 (7)	0.773 (4)	0.612 (11)	0.688 (8)	0.729 (5)	**0.841 (1)**
WPBC (arff)	0.21 (12)	0.224 (8)	0.23 (6)	0.231 (5)	0.233 (4)	0.223 (9)	0.225 (7)	**0.248 (1)**	0.223 (9)	0.223 (9)	0.245 (2)	**0.244 (3)**
AVG	0.351 (12)	0.462 (8)	0.509 (3)	0.538 (2)	0.447 (9)	0.436 (10)	0.378 (11)	0.466 (7)	0.49 (6)	0.503 (4)	0.503 (4)	**0.565 (1)**

Сравнение с другими методами (время работы)

TABLE 7: Run time (in seconds) of detectors (average of 10 independent trials, the fastest is highlighted in bold); rank is shown in parenthesis (lower is better). ECOD is one of the fastest algorithms.

Data	ABOD	CBLOF	HBOS	IForest	KNN	LODA	LOF	LSCP	OCSVM	PCA	SUOD	ECOD
Arrhythmia (mat)	0.236 (7)	0.258 (8)	**0.201 (6)**	**0.306 (10)**	0.077 (5)	0.05 (2)	0.066 (4)	0.984 (11)	**0.042 (1)**	0.058 (3)	2.004 (12)	0.266 (9)
Breastw (mat)	0.17 (9)	0.076 (8)	0.004 (2)	0.384 (10)	0.038 (7)	0.037 (6)	0.007 (3)	0.524 (11)	0.01 (4)	**0.002 (1)**	1.431 (12)	**0.024 (5)**
Cardio (mat)	0.546 (10)	0.175 (8)	0.011 (2)	0.426 (9)	0.158 (7)	0.06 (4)	0.116 (6)	1.512 (11)	0.086 (5)	**0.005 (1)**	1.811 (12)	0.053 (3)
Ionosphere (mat)	0.083 (9)	0.053 (8)	0.012 (4)	0.271 (10)	**0.015 (5)**	0.025 (6)	0.006 (3)	0.45 (11)	0.004 (2)	**0.003 (1)**	1.458 (12)	0.047 (7)
Lympho (mat)	0.036 (8)	0.051 (9)	**0.009 (5)**	**0.253 (10)**	0.008 (4)	0.023 (6)	0.003 (3)	0.32 (11)	**0.001 (1)**	0.002 (2)	1.399 (12)	0.036 (8)
Mammography (mat)	1.701 (10)	0.268 (6)	**0.006 (1)**	0.689 (8)	0.473 (7)	0.091 (4)	0.255 (5)	4.6 (12)	1.364 (9)	0.007 (2)	2.517 (11)	**0.044 (3)**
Optdigits (mat)	2.154 (10)	0.433 (5)	**0.032 (1)**	0.725 (6)	1.457 (9)	0.063 (3)	1.371 (8)	14.566 (12)	1.169 (7)	0.046 (2)	7.114 (11)	0.229 (4)
Pima (mat)	0.176 (9)	0.072 (8)	**0.001 (1)**	0.267 (10)	**0.03 (7)**	0.023 (6)	0.009 (4)	0.577 (11)	0.007 (3)	0.003 (2)	1.452 (12)	0.023 (6)
Satellite (mat)	1.602 (10)	**0.399 (5)**	**0.018 (1)**	0.669 (6)	0.867 (8)	0.071 (3)	0.835 (7)	8.549 (12)	1.075 (9)	0.024 (2)	5.173 (11)	0.158 (4)
Satimage-2 (mat)	1.43 (10)	**0.354 (5)**	0.017 (2)	0.546 (6)	0.667 (8)	0.063 (3)	0.64 (7)	7.699 (12)	0.91 (9)	0.017 (2)	4.742 (11)	0.146 (4)
Shuttle (mat)	2.01 (10)	0.175 (4)	0.006 (2)	**0.634 (6)**	0.788 (8)	0.076 (3)	0.728 (7)	7.973 (12)	1.33 (9)	**0.003 (1)**	3.566 (11)	**0.189 (5)**
Speech (mat)	**7.85 (10)**	2.036 (5)	0.167 (2)	3.589 (6)	7.842 (9)	**0.154 (1)**	7.274 (8)	49.828 (12)	4.272 (7)	1.052 (3)	14.31 (11)	1.148 (4)
Wbc (mat)	0.08 (9)	0.06 (8)	0.008 (4)	0.239 (10)	0.015 (5)	0.021 (6)	0.007 (3)	0.47 (11)	0.004 (2)	**0.002 (1)**	1.495 (12)	**0.055 (7)**
Wine (mat)	0.026 (8)	0.046 (9)	0.005 (5)	0.243 (10)	0.005 (5)	0.02 (7)	0.002 (3)	0.313 (11)	**0.001 (1)**	0.002 (3)	1.421 (12)	**0.019 (6)**
Arrhythmia (arff)	0.226 (8)	0.23 (9)	0.169 (6)	0.249 (10)	0.065 (4)	0.043 (2)	0.057 (3)	0.947 (11)	**0.039 (1)**	0.095 (5)	**2.025 (12)**	**0.226 (8)**
Cardiotocography (arff)	0.368 (10)	0.124 (7)	0.007 (2)	0.286 (9)	0.138 (8)	0.049 (4)	0.099 (6)	1.769 (11)	0.084 (5)	**0.007 (2)**	1.876 (12)	0.037 (3)
HeartDisease (arff)	0.048 (9)	0.044 (8)	**0.005 (4)**	0.21 (10)	0.008 (5)	0.018 (7)	0.004 (3)	0.366 (11)	**0.001 (1)**	0.002 (2)	1.38 (12)	**0.016 (6)**
Hepatitis (arff)	0.013 (6)	0.028 (9)	0.004 (5)	0.187 (10)	0.003 (4)	0.016 (7)	0.002 (3)	0.292 (11)	0.001 (2)	0.001 (2)	1.378 (12)	**0.021 (8)**
InternetAds (arff)	5.648 (9)	1.36 (3)	**0.493 (2)**	5.605 (8)	5.252 (6)	**0.227 (1)**	6.11 (10)	58.435 (12)	4.159 (5)	5.496 (7)	21.974 (11)	1.855 (4)
Ionosphere (arff)	**0.075 (9)**	0.049 (8)	0.012 (4)	0.236 (10)	**0.014 (5)**	0.02 (6)	0.007 (3)	0.447 (11)	0.004 (2)	0.004 (2)	1.498 (12)	0.036 (7)
KDDCup99 (arff)	136.469 (9)	**2.1 (5)**	**0.135 (1)**	**7.714 (6)**	133.6 (8)	0.409 (3)	141.9 (10)	839.8 (12)	175.1 (11)	**0.301 (2)**	68.726 (7)	**0.965 (4)**
Lymphography (arff)	0.039 (8)	0.053 (9)	0.007 (5)	**0.255 (10)**	0.007 (5)	0.023 (7)	0.003 (3)	0.31 (11)	**0.001 (1)**	0.002 (2)	1.503 (12)	**0.02 (6)**
Pima (arff)	0.201 (9)	0.087 (8)	0.004 (2)	0.319 (10)	**0.034 (6)**	0.029 (5)	0.014 (4)	0.573 (11)	0.013 (3)	**0.002 (1)**	1.43 (12)	**0.043 (7)**
Shuttle (arff)	0.248 (8)	0.079 (7)	0.004 (2)	0.366 (9)	0.05 (6)	0.033 (5)	**0.023 (4)**	0.712 (10)	0.018 (3)	**0.003 (1)**	1.481 (12)	0.912 (11)
SpamBase (arff)	1.329 (10)	0.303 (5)	**0.017 (1)**	0.514 (6)	0.904 (9)	0.061 (3)	0.888 (8)	8.349 (12)	0.666 (7)	0.032 (2)	3.836 (11)	**0.109 (4)**
Stamps (arff)	0.071 (9)	0.055 (8)	0.004 (4)	0.233 (10)	0.013 (5)	0.018 (7)	0.003 (3)	0.384 (11)	0.002 (2)	0.002 (2)	1.382 (12)	**0.014 (6)**
Waveform (arff)	1.018 (10)	0.244 (5)	0.01 (2)	0.478 (8)	**0.521 (9)**	0.072 (4)	0.429 (7)	3.479 (12)	0.261 (6)	**0.008 (1)**	2.357 (11)	**0.047 (3)**
WBC (arff)	0.048 (8)	0.05 (9)	0.004 (4)	**0.243 (10)**	0.009 (5)	0.021 (7)	0.003 (3)	0.335 (11)	0.001 (2)	0.001 (2)	1.367 (12)	0.013 (6)
WDBC (arff)	0.092 (9)	0.062 (8)	0.009 (4)	0.245 (10)	0.016 (5)	0.022 (6)	0.007 (3)	0.466 (11)	0.005 (2)	**0.004 (1)**	1.5 (12)	**0.032 (7)**
WPBC (arff)	0.046 (7)	0.052 (8)	0.01 (5)	0.237 (10)	0.008 (4)	0.021 (6)	0.004 (3)	**0.353 (11)**	**0.002 (1)**	0.003 (2)	1.383 (12)	**0.062 (9)**
AVG	5.468 (9)	0.312 (5)	**0.046 (1)**	0.887 (6)	5.105 (7)	0.062 (2)	5.362 (8)	33.847 (12)	6.355 (11)	0.24 (4)	5.5 (10)	**0.228 (3)**