FACULTY
OF INFORMATICS
Masaryk University

RF-OEX

Outlier factor and anomaly detection

Petr Matonoha
485185@mail.muni.cz

April 20, 2021

Table of Contents

FACULTY
OF INFORMATICS
Masaryk University

Random forest

A random forest is a group of decision trees, which make an independent decision, then the random forest makes the decision based on a majority vote.
The main advantage of the random forest is that it does not overfit.

FACULTY
OF INFORMATICS
Masaryk University

Random forest classifier

A random forest classifier is composed of tree-structured classifiers $h(x, \Theta_k), k = 1, 2, \ldots$, where the Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x

FACULTY
OF INFORMATICS
Masaryk University

RF-OEX

Random Forest-outlier explanation (RF-OEX) is an outlier detection method focused on class-based anomalies.
The main idea of RF-OEX is using proximity matrix differently than classical random forest.
The main difference is how RF-OEX exploits information about the class label. The outlier factor of instance p is calculated as the sum of three different values.

FACULTY
OF INFORMATICS
Masaryk University

Proximity matrix

Proximity value between element *a* and *b* is the number of trees that classified *a* as the same class as *b* divided by the total number of trees.
The proximity matrix contains proximity values between all elements.

FACULTY
OF INFORMATICS
Masaryk University

Outlier factor

$$OF(p) = OF_1(p)_{same-class} + OF_2(p)_{misclassification} + OF_3(p)_{ambiguity}$$

FACULTY
OF INFORMATICS
Masaryk University

OF_1

$OF_1(p)_{same-class}$:
$OF_1(p) = \frac{1}{\sum_{cl(p)=cl(q)} Prox(p,q)}$, where
$cl(p)$ is the class of the point p.
The result have to be normalized, because classes are not uniformly distributed.

FACULTY
OF INFORMATICS
Masaryk University

OF_2

$OF_2(p)_{misclassification}$:
We define $Top_{C_p}(p)$ as $|C_p|$ (all points with class $cl(p)$) points, which are the closest to p.
To be comparable with OF_1 and OF_3, we multiply the value with the constant $c = \frac{max_{q \in DB} OF_1(q)}{4}$,

$$OF_2(p) = c \cdot \frac{|\{q \mid cl(q) \neq cl(p) \ \&\ q \in Top_{|C_p|}(p)\}|}{|C_p|}$$

FACULTY
OF INFORMATICS
Masaryk University

OF_3

$OF_3(p)_{ambiguity}$:
For increasing importance of outliers, which are distant from other points.

$$OF_3(p) = c \cdot \frac{|C_p| - \sum_{q \in Top_{|C_p|}(p)} Prox(p,q)}{|C_p|}$$

FACULTY
OF INFORMATICS
Masaryk University

Setting

Classifiers: J48, JRip, PART, SMO
Removed: 1, 3, 5 % of outliers
Datasets: We are expecting improvement in all used datasets, with one exception *vehicle* dataset.

FACULTY
OF INFORMATICS
Masaryk University

Different weights

Testing $OF(p) = A \cdot OF_1(p) + B \cdot OF_2(p) + C \cdot OF_3(p)$ and $OF(p) = OF_1(p) \cdot OF_3(p)$

FACULTY
OF INFORMATICS
Masaryk University

dataset	111	110	001	101-multi
analcatdata_dmft	19.16	19.25	19.2	19.47
balance-scale	82.11	82.41	82.93	82.16
blood-transfusion-service	77.47	77.42	77.64	77.63
cmc	51.2	51.42	51.58	51.11
dresses-sales	57.9	57.93	59.33	57.9
eucalyptus	55.47	54.84	55.92	54.88
ilpd	68.98	68.78	69.8	68.87
pima-diabetes	75.49	75.51	75.04	75.49
qsar-biodeg	83.7	83.51	83.93	83.98
vehicle	70.69	70.78	71.1	70.73
comparison with 111	0	0	0.4304	0.0024

FACULTY
OF INFORMATICS
Masaryk University

dataset	111	001	102	103	104	112
analcatdata_dmft	19.44	19.48	19.25	19.36	19.47	19.25
balance-scale	82.23	82.45	82.04	82.44	82.35	81.84
blood-transfusion-service	77.53	77.83	77.75	77.75	77.71	77.63
cmc	51.1	51.09	51.4	51.27	51.52	51.4
dresses-sales	58.75	58.28	58.65	58.55	58.38	58.65
eucalyptus	54.96	56.05	55.05	55.61	55.37	54.92
ilpd	69.21	69.44	69.25	69.54	69.67	69.34
pima-diabetes	75.26	75.69	75.4	75.67	75.51	75.42
qsar-biodeg	83.97	84.03	83.82	83.97	84.06	83.97
vehicle	70.93	70.49	70.44	70.77	70.67	70.44
comparison with 111	0	0.146	-0.032	0.155	0.132	-0.051

FACULTY
OF INFORMATICS
Masaryk University

Ground truth data

Testing individual OF_i on ground truth data

FACULTY
OF INFORMATICS
Masaryk University

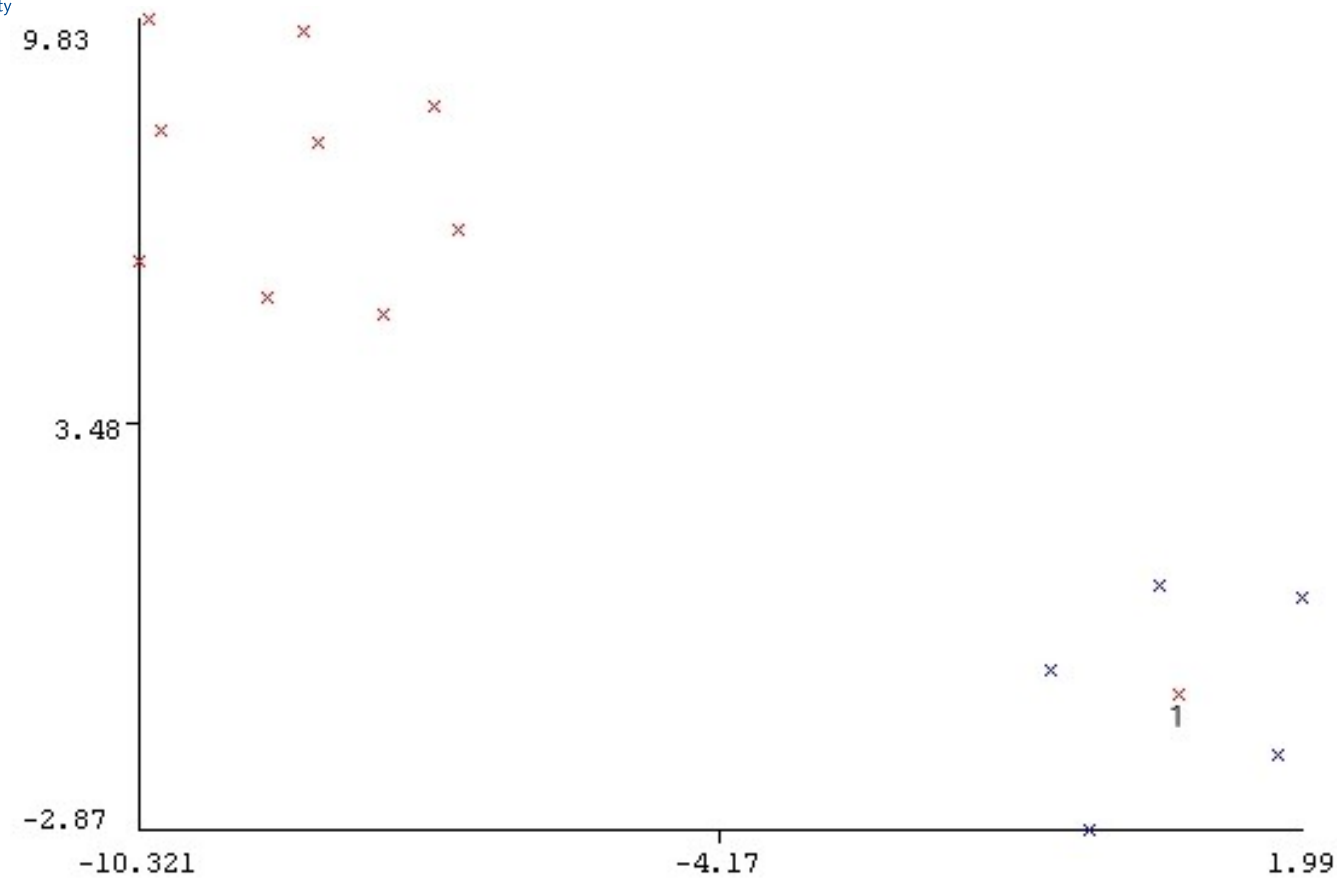

Figure: PositivesNegativesAroundPositive

Anomaly 1: OF_3: 5.62 % OF_1 94.37 %

FACULTY
OF INFORMATICS
Masaryk University

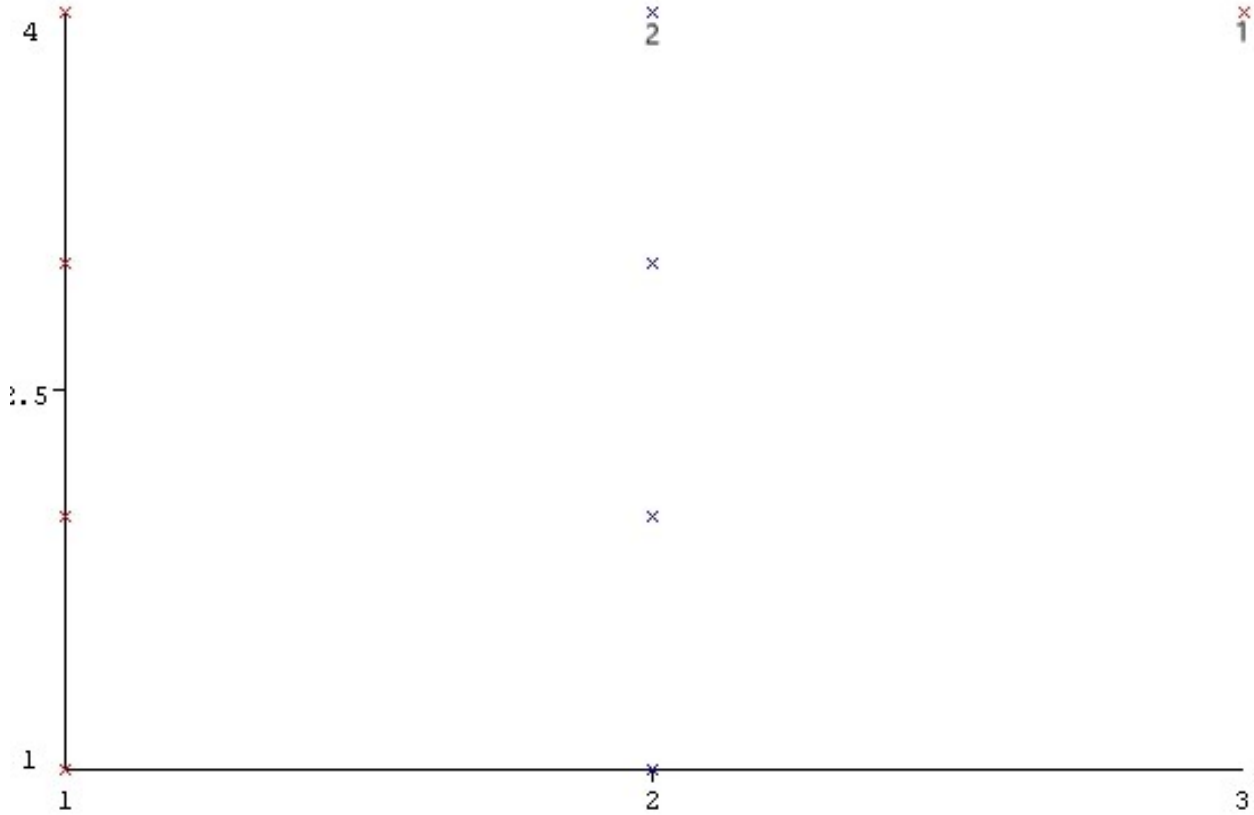

Figure: Rastr-3x4-0

Anomaly 1: OF_3: 9.05 % OF_2 4.97 % OF_1 85.96 %
Anomaly 2: OF_3: 4.9 % OF_1 95.09 %

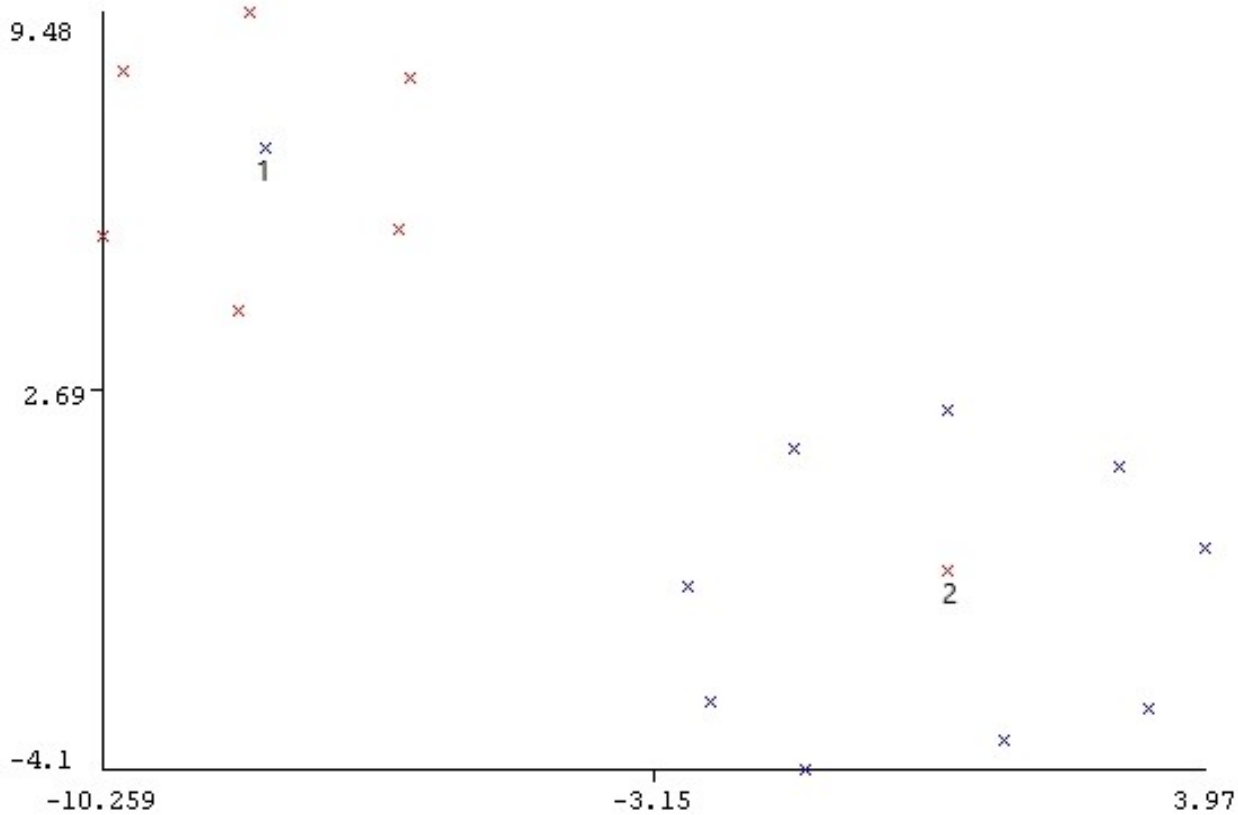

Figure: TwoCircles

Anomaly 1: OF_3: 7.44 % OF_2 2.08 % OF_1 90.47 %
Anomaly 2: OF_3: 8.27 % OF_2 3.39 % OF_1 88.33 %

FACULTY
OF INFORMATICS
Masaryk University

Influence OF_i on OF

Testing influence of OF_i on used datasets.

$$\text{Influence } OF_i = \frac{\sum_{q \in DB} \frac{|OF_i(q)|}{|OF_1(q)|+|OF_2(q)|+|OF_3(q)|}}{|DB|}$$

FACULTY
OF INFORMATICS
Masaryk University

5 % elements with the highest outlier factor

dataset	OF_1 %	OF_2 %	OF_3 %
analcatdata_dmft	73.08	0.08	26.82
balance-scale	51.6	2.65	45.74
blood-transfusion-service	81.85	1.96	16.17
dresses-sales	88.34	0.25	11.39
eucalyptus	65.69	0.38	33.92
ilpd	84.39	0.46	15.13
pima-diabetes	86.93	0.39	12.66
qsar-biodeg	94.1	0.0	5.89
average	78.25	0.77	20.97

FACULTY
OF INFORMATICS
Masaryk University

All elements

dataset	OF_1 %	OF_2 %	OF_3 %
analcatdata_dmft	56.42	0.0	43.56
balance-scale	50.21	0.32	49.45
blood-transfusion-service	66.5	0.7	32.78
dresses-sales	67.71	0.02	32.25
eucalyptus	42.76	0.08	57.15
ilpd	67.93	0.03	32.02
pima-diabetes	73.37	0.05	26.57
qsar-biodeg	85.9	0.0	14.09
average	63.85	0.15	35.98

FACULTY
OF INFORMATICS
Masaryk University

Results

The highest influence of OF_2 for 5 % elements with the highest outlier factor: balance-scale 2.65 %
The highest influence of OF_2 for all elements:
blood-transfusion-service 0.7 %

FACULTY
OF INFORMATICS
Masaryk University

Linear regression

Our problem is choosing the best weights for OF.
To find them we decided to use linear regression.
Influence of OF_2 is low in all shown datasets, hence we decided to not use it in linear regression
For linear regression were used weights:
$101, 100, 001, 102, 201, 103, 301, 104, 401, 203, 302$
We used basic model sklearn.linear_model.LinearRegression.
We were trying to find coefficients function, such that
$gain\_clfBest = A \cdot OF_1 + B \cdot OF_3 + C$
We were using leave one out method, to be able to test results.

FACULTY
OF INFORMATICS
Masaryk University

Linear regression with all weights

without dataset	score	A	B
analcatdata_dmft	0.3994	-0.00078	-0.00002
balance-scale	0.47336	-0.00075	-0.00004
blood-transfusion-service	0.47385	-0.00077	-0.00005
dresses-sales	0.4029	-0.0006	0.00004
eucalyptus	0.42369	-0.00049	-0.00003
ilpd	0.54039	-0.00058	0.00015
pima-diabetes	0.47233	-0.00064	-0.00002
qsar-biodeg	0.41695	-0.00071	-0.00002

FACULTY
OF INFORMATICS
Masaryk University

Linear regression with six the best performing weights

without dataset	score	A	B
analcatdata_dmft	0.53131	-0.00112	-0.00009
balance-scale	0.7033	-0.00192	0.00011
blood-transfusion-service	0.78788	-0.00197	0.00007
dresses-sales	0.51715	-0.00055	-0.00007
eucalyptus	0.4756	-0.00104	-0.00001
ilpd	0.77981	-0.00139	0.00024
pima-diabetes	0.76615	-0.00172	0.00011
qsar-biodeg	0.76329	-0.00209	0.00015

FACULTY
OF INFORMATICS
Masaryk University

Random forest
RF-OEX
Experiments

FACULTY
OF INFORMATICS
Masaryk University

Thank you for your attention!