J. Yaghob (Ed.): ITAT 2015 pp. 143–149
Charles University in Prague, Prague, 2015

Evaluation of Association Rules Extracted during Anomaly Explanation

Martin Kopp[1,2,3] Martin Holeňa[1,3]

[1] Faculty of Information Technology, Czech Technical University in Prague
Thákurova 9, 160 00 Prague
[2] Cisco Systems, Cognitive Research Team in Prague
[3] Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07 Prague

Abstract: Discovering anomalies within data is nowadays very important, because it helps to uncover interesting events. Consequently, a considerable amount of anomaly detection algorithms was proposed in the last few years. Only a few papers about anomaly detection at least mentioned why some samples were labelled as anomalous. Therefore, we proposed a method allowing to extract rules explaining the anomaly from an ensemble of specifically trained decision trees, called sapling random forest.

Our method is able to interpret the output of an arbitrary anomaly detector. The explanation is given as conjunctions of atomic conditions, which can be viewed as antecedents of association rules. In this work we focus on selection, post processing and evaluation of those rules. The main goal is to present a small number of the most important rules. To achieve this, we use quality measures such as lift and confidence boost. The resulting sets of rules are experimentally and empirically evaluated on two artificial datasets and one real-world dataset.

Keywords: Anomaly detection, anomaly interpretation, association rules, confidence boost, random forest

1 Introduction

According to an IBM research [13] there were 2,7 zettabytes of data in the digital universe at April 2012 and this amount is doubling approximately every 40 months.

Not only it is almost impossible to process such huge amounts of data, we are actually not interested in the raw data, but rather in the salient knowledge and interesting patterns contained in them. This is the reason why anomaly detection, especially unsupervised anomaly detection, becomes more and more important [1, 23]. Despite it can be formalised as a binary classification, it entails different issues and challenges than those in supervised classification. For example, anomalous events often adapt to appear normally and even normal behaviour evolve over time. Furthermore, defining a normal regions is very difficult, especially when the boundary between normal and anomalous is not always precise.

For the purposes of this paper consider anomalies equal to outliers as defined by Hawkins [11]: " *An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.*"

The more formal definition would necessarily reduce the amount of plausible anomaly detectors and/or application domains. This is in conflict with our goal to provide a solution as general as possible.

Even though anomaly detection techniques are aimed at only a minority of samples, the importance and demand for them grows rapidly. The real world applications range from the network security [10], bioinformatics [24] or financial fraud detection [22] to the astronomy and space exploration [9].

The identification of anomalies is only a half of the whole task. The second and equally important half is the interpretation. In high dimensional domains, like the network security or bioinformatics, where hundreds or even thousands of features are common, the proper interpretation is crucial.Therefore, anomalies have to be interpreted clearly, as a feature subset that explains its deviation from ordinary data, or even better as a set of association rules.

In [21] we proposed method of anomaly explanation based upon specifically trained ensembles of decision trees called sapling random forest (SRF). The main idea behind it is to view the explanation as a feature selection and classification problem. Specifically, the goal is to find features in which the margin betweenı anomalous sample and the normal samples is maximised. Therefore, SRF returns subset of features, respectively rules on these features describing why this sample has been identified as an anomaly.

The main drawback of the direct rule extraction from our sapling random forests is the big number of rules with some of them introduced by unfortunate training set selection. Partially, these issues can be solved by confidence and / or support thresholds. But for our ultimate goal to present the minimal number of rules containing the maximal amount of useful information, such a simple approach is insufficient. Therefore, in this paper we focus on proper selection, post processing and evaluation of rulesets extracted from sapling random forests during anomaly explanation. We tested association rules quality measures such as lift and confidence boost. This paper is work in progress and we would like to extend the number of tested quality measures by some subjective measures like novelty.

The rest of this paper is organised as follows. The next section briefly reviews related work. Section 3 describes the SRF principles and its training followed by the rule extraction process in Section 4. The selected quality mea-

sures of association rules are presented in Section 5. Experimental evaluation is described in Section 6 and Section 7 concludes the paper.

2 Related Work

For more information about the anomaly detection we refer to the recent book of Aggarwall [1]. This book provides an exhaustive listing of anomaly detection algorithms and their applications in different domains. Another source may be [5], which is briefer but very well written. To our best knowledge, there have been only few works addressing not only identification of anomalies, but also their explanation.

Knorr et al. [15] focused on what kind of knowledge should be extracted and provided to the user. Strong and weak outliers were defined and searched within data by distance-based algorithms described in detail in [14].

Dang et al. [7] presented an algorithm identifying and explaining anomalies. The algorithm starts by selecting a set of neighbouring samples based on quadratic entropy that are presented to a Fisher linear discriminant classifier to seek for an optimal half-space, in which a detected anomaly is well separated. The process of interpretation is entangled with the presented method of identification of anomalies. The difference to our work is that SRF can be used after an arbitrary anomaly detection algorithm to interpret its results.

The most similar to our approach and most recent is [20]. Their approach, as well as ours, can interpret output of an arbitrary anomaly detector as a subset of features. They use classification accuracy for outlier ranking. The main drawback of this approach is that it needs balanced training sets which are created by sampling artificial samples around the anomalous point. With respect to this work, our approach can handle unbalanced training sets easily and returns not only feature subsets but feature subsets with rules on them, providing even more information about the anomaly. Furthermore, we simplify the analysis by clustering, which enables to interpret similar anomalies at once [16].

On the other hand, there are many papers about association rules. This paper was inspired mainly by [3], which is about measuring redundancy and information quality of sets of association rules. The author presents a measure called confidence boost and an algorithm to produce a small set of association rules using this measure. A really extensive list of interestingness measures can be found in [12]. There is a lot of inspiration for our future work.

An alternative approach, well described in [6], may be so called subjective measures. A typical example is the novelty, sometimes called unexpectedness, of a rule with respect to user provided domain knowledge or against the another rule set. Because these terms are ambiguous there are multiple approaches of measuring them. An approach in [18] inspired us for our future work.

3 Sapling Random Forest

This section outlines principles of sapling random forests. SRF is a method able to explain an output of an arbitrary anomaly detector, proposed by us in [21]. It is a random forest of specifically trained decision trees. Because produced trees are small they are called saplings rather than trees. Produced explanations show features in which inspected samples differ the most from the rest of data. These features are used to produce association rules, which are more informative than only a set of features. An outline of the whole method is at Algorithm 1.

Algorithm 1 Algorithm summary

$y \leftarrow anomalyDetector(data)$
for all data(y ==anomaly) **do**
  $T \leftarrow createTrainingSet(size, method)$
  $t \leftarrow trainTree(T)$
  $SRF \leftarrow t$
end for
$extractRules(SRF)$

3.1 Training Set Selection

Dataset $\boldsymbol{\mathcal{X}} = \{x_1, x_2, \ldots, x_l\}$, where $x \in \mathbb{R}^d$, can be split into two disjoint sets $\boldsymbol{\mathcal{X}}^a$, containing anomalous samples, and $\boldsymbol{\mathcal{X}}^n$,containing normal samples. Then, a training set $\boldsymbol{\mathcal{T}}$ contains the anomaly x^a as one class and a subset of $\boldsymbol{\mathcal{X}}^n$ as the other. The first strategy of creating training sets is to select k nearest neighbours of x^a from $\boldsymbol{\mathcal{X}}^n$. This strategy is sensible for algorithms detecting local anomalies, as according to [8] they are more general than algorithms detecting global anomalies. The drawback of this strategy is a computational complexity.

The second strategy is to select k samples randomly from $\boldsymbol{\mathcal{X}}^n$ with uniform probability. The advantage of this approach is a possibility to generate more than one training set per anomaly by repeating the sampling process. More training sets lead to more saplings per anomaly and to more robust explanation, but at the expense of the more complicated aggregation of rules extracted from them (see Section 4). A comparison of both approaches can be found in [21].

3.2 Training a Sapling

For simplicity consider sapling a binary decision tree with typical height 1-3. In the SRF method, there are always two leaves at the maximal depth, one of which contains only an anomaly x^a and the other containing only normal samples. The saplings small height has two reasons. First, training sets are relatively small. Second, according to the anomaly isolation approach [19], if the analysed sample is an anomaly, it should be separated easily from the rest of data, resulting again into small trees. Therefore, if the

height of a sapling is higher than expected it should be taken into consideration that the explained sample may not be an anomaly.

The standard procedure, to find the splitting function h for a new internal node, is maximising an information gain over the space of all possible splitting functions $\mathcal{H}$ as

$$\arg\max_{h\in\mathcal{H}} - \sum_{b\in\{L,R\}} \frac{|\mathcal{S}^b(h)|}{|\mathcal{S}|} H(\mathcal{S}^b), \tag{1}$$

where $\mathcal{S}$ is the subset of the training set $\mathcal{T}$ reaching the leaf being split, $\mathcal{S}^L(h) = \{x \in \mathcal{S} | h(x) = +1\}$ and $\mathcal{S}^R(h) = \{x \in \mathcal{S} | h(x) = -1\}$ and $H(\mathcal{S})$ is an entropy of $\mathcal{S}$.

The second commonly used approach involves minimising the Gini impurity.

$$\arg\min_{h\in\mathcal{H}} \sum_{b\in\{L,R\}} 1 - \left(\frac{|x^a|}{(\mathcal{S}^b(h))}\right)^2 - \left(\frac{|x^n|}{(\mathcal{S}^b(h))}\right)^2 \tag{2}$$

For experiments presented in this paper we used information gain.

4 Extraction and Evaluation of Rules

Once a sapling is grown, it is used to explain the anomaly x^a. Let $h_{j_1,\theta_1},\ldots,h_{j_d,\theta_d}$ be the set of splitting functions, with features $j_1,\ldots,j_d$ and threshold $\theta_1,\ldots,\theta_d$, used in the inner nodes on the path from the root to the leaf with the anomaly x^a. Then x^a is explained as a conjunction of atomic conditions :

$$c = (x_{j_1} > \theta_1) \wedge (x_{j_2} > \theta_2) \wedge \ldots \wedge (x_{j_d} > \theta_d), \tag{3}$$

which is the output of the algorithm. This conjunction can be read as "the sample is anomalous because it is greater than threshold θ_1 in feature j_1 and greater than θ_2 in feature j_2 and ... than majority (or nearest neighbour) samples." Because resulting trees are very small, the explanation is compact.

The situation is more difficult, when more saplings per anomaly have been grown, as each sapling provides one conjunction of type (3). Using more than one sapling per anomaly improves robustness for training sets created by uniform sampling. The problem is that returning set of all conjunctions $\mathcal{C}$ is undesirable, as the primary objective — explanation of the anomaly to a human — would not be met. Hence, the algorithm needs to aggregate conjunctions in $\mathcal{C}$.

For simplicity of the following notation consider $2d$ items, in such a way that 2 items are assigned to each feature, one for "<" rules, the other for ">" rules. Denote this $2d$ set of items $\mathcal{F}$. Then we can group rules into the rule sets $\mathcal{R}_f$ according to the item set $f \subseteq \mathcal{F}$ they share.

Based on $|\mathcal{R}_f|$ the algorithm discards groups of low importance by sorting them in descend order, and then using only the first k groups such, that their cumulative frequency is greater than a threshold τ, which we recommend to be 0.90 or 0.95. Using the adopted notation, k is determined as

$$k = \arg\min_k \frac{1}{\sum_{f\in\mathcal{F}} |\mathcal{R}_f|} \sum_{i=1}^{k} |\mathcal{R}_{f_i}| > \tau, \tag{4}$$

where it is assumed, that $\mathcal{R}_f$ are sorted according to their size to simplify the notation. We have also investigated the complementary approach, where groups are selected, if they were used with a frequency higher than a specified threshold. But the presented strategy based on the cumulative frequency showed more consistent results in our experiments.

Once the set of groups with decision rules is selected, we create one rule $\bar{r}_f$ for every rule set $\mathcal{R}_f$.Thresholds for each item f_j are calculated as an average of all thresholds within the rule set $\mathcal{R}_f$.

$$\bar{\theta}_j = \frac{1}{|\mathcal{R}_f|} \sum_{i=1}^{|\mathcal{R}_{f_i}|} \theta_{i,j} \tag{5}$$

By this approach we obtain one representative rule for each feature set f as:

$$\bar{c_f} = x_{j_1} > \bar{\theta}_1 \wedge x_{j_2} > \bar{\theta}_2 \wedge \ldots \wedge x_{j_t} > \bar{\theta}_t. \tag{6}$$

5 Measuring Quality of Rules

This section reviews selected quality measures of association rules. Typical association rules are in the form $\mathcal{A} \to \mathcal{Y}$, where $\mathcal{A}, \mathcal{Y}$ are item sets. In rules extracted from SRF, items are atomic conditions and $\mathcal{Y}$ always means: "is anomalous". Therefore, our rules are in the form:

$$r_f = c_f \to y, \tag{7}$$

where c is a conjunction of atomic conditions like (3), $y = x \in \mathcal{X}^a$ and $f \subseteq \mathcal{F}$. The r_f in its full form then look as:

$$r_f(x) = x_{f_1} > \theta_{f_1} \wedge x_{f_2} > \theta_{f_2} \wedge \ldots \wedge x_{f_n} > \theta_{f_n} \to x \in \mathcal{X}^a, \tag{8}$$

where n is a maximal index in the itemset f.

For this kind of rules support [2] is calculated as:

$$\text{supp}(c_f) = \frac{|\{c_f(x) | x \in \mathcal{X}\}|}{|\mathcal{X}|}, \tag{9}$$

$$\text{supp}(y) = \frac{|\mathcal{X}^a|}{|\mathcal{X}|}. \tag{10}$$

and gives the proportion of data points which satisfy the antecedent c, respectively the consequent y. It is used to measure the importance of a rule or as a frequency constrain. The disadvantage of support is that infrequent rules are often discarded. This is much bigger problem than it could seem because we are generating rules for anomalies, which are rare by definition.

Another frequently used measure is confidence [2]:

$$\mathrm{conf}(c_f \to y) = \frac{\mathrm{supp}(c_f \to y)}{\mathrm{supp}(c_f)}. \tag{11}$$

It estimates the conditional probability of the consequent being true on condition that the antecendent is true. The trouble with confidence is caused by its sensitivity to the frequency of y. Because all rules extracted from SRF have the same consequent the rule ranking produced by lift a confidence would be the same.

The third measure we used is lift [4]:

$$\mathrm{lift}(c_f \to y) = \frac{\mathrm{conf}(c_f \to y)}{\mathrm{supp}(y)}, \tag{12}$$

which measures how many times more often the antecedent c and consequent y occur together than expected if they were statistically independent. Lift does not suffer from the rare items problem. Because in our experiments the consequent will always the same frequency there is no need to measure both

Finally, confidence boost introduced by Balcázar [3] is calculated as:

$$\beta(r_f) = \frac{\mathrm{conf}(r_f)}{\max\{\mathrm{conf}(r'_{f'}) | \mathrm{supp}(r'_{f'}) > \sigma, r_f \not\equiv r'_{f'}, f' \subseteq f\}}, \tag{13}$$

where σ is support threshold and $r_f \not\equiv r'_{f'}$ denotes the inequivalence of rules r_f and $r'_{f'}$, which for our simple case where all consequents are the same, means that $f \neq f'$. From (13) is evident that $f' \subset f$.

If the set of confidences in denominator is empty, the confidence boost is by convention set to infinity.

6 Experiments

For the experimental evaluation we used the synthetical three layer donut, the well known Fisher's iris and the Letter recognition data set from the UCI repository [17].

6.1 Three Layer Donut

The three layer donut dataset contains 1000 normal samples forming the two dimensional toroid (donut). There are 200 anomalies, one half inside the toroid and the second half out of it. For this dataset we created 10 rules per anomaly, using SRF, resulting in 2000 rules. After the simple aggregation, described in Section 4, only 8 rules left. All of them printed in Table 1, sorted by their respective support. All rules before aggregation were used to calculate confidence boost. Otherwise, many rules would have confidence boost equal to infinity because there are no rules defined on any subset of their item sets f.

The fact is that rules $r_5 - r_8$ have quite small support but, according to the other measures and our intuition, they are very important. The small supports is due to the small number of data points explained, lets recall that anomalies are only one sixth of all data points in this dataset. Both, lift and confidence boost, mostly reflects our subjective expectations.

rule	supp	lift	β
$r_1 = x_1 > -0.33 \wedge x_2 > -0.39$	0.38	1.64	0.27
$r_2 = x_1 > -0.33 \wedge x_2 < 0.3$	0.37	1.60	0.27
$r_3 = x_1 < 0.4 \wedge x_2 < 0.34$	0.37	1.49	0.25
$r_4 = x_1 < 0.37 \wedge x_2 > -0.37$	0.37	1.57	0.26
$r_5 = x_2 > 2.2$	0.02	6.00	1.00
$r_6 = x_1 > 2.3$	0.02	6.00	1.00
$r_7 = x_2 < -2.4$	0.01	6.00	1.00
$r_8 = x_1 < -2.4$	0.01	6.00	1.00

Table 1: Aggregated rules with their quality measures for the three layer donut dataset, sorted by their respective supports.

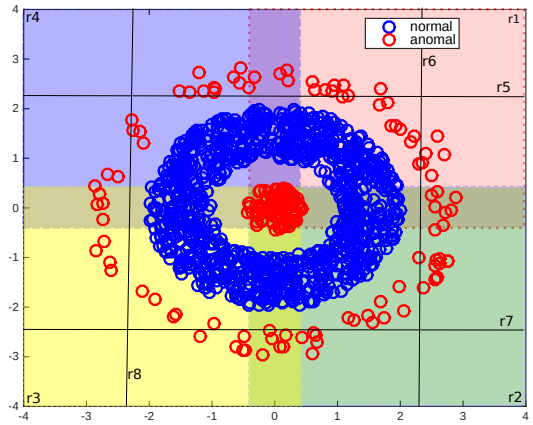

Figure 1: Three layer donut dataset with plotted rules $r_1 - r_8$. Rules $r_1 - r_4$ are plotted as filled squares and rules $r_5 - r_8$ as half-planes delimited by solid lines.

All rules are depicted at Figure 1. Its evident that presented rules cannot separate anomalies from normal samples perfectly. Especially difficult are anomalies inside the donut. To separate those inner anomalies perfectly it would be necessary to combine more rules together, for example r_1 and r_3 or r_2 and r_4.

6.2 Iris

The virginica species were selected as anomalous class for the iris data set. Five rules per anomaly were produced using SRF, resulting in 250 rules. After aggregation we have got 6 rules. They are written in Table 2 with their respective quality measures. Confidence boost was calculated using all 250 rules.

The main problem with all those rules is that almost every one of them can sufficiently separate anomalies from

rule	supp	lift	β
$r_1 = x_1 > 6 \wedge x_4 > 1.7$	0.25	3.00	1.00
$r_2 = x4 > 2$	0.19	3.00	1.00
$r_3 = x_3 > 5.5$	0.17	3.00	1.00
$r_4 = x_2 < 2.8 \wedge x_4 > 1.6$	0.06	3.00	1.00
$r_5 = x_1 > 7.3$	0.05	3.00	1.00
$r_6 = x_2 < 2.2$	0.03	0.75	0.25

Table 2: Aggregated rules with their quality measures for the iris dataset , sorted by their respective supports.

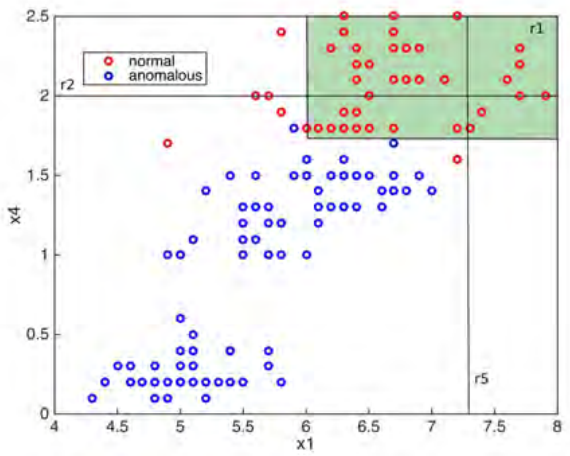

Figure 2: The iris dataset with r_1 plotted as a filled rectangle, and rules r_2 and r_5 as half-planes delimited by solid lines.

normal samples. No one of presented measures could help in selecting the most informative, yet small as possible, set of rules. Because presented rules are seen informationally equivalent by all quality measures. This doesn't say much about difference between the quality measures but it justifies the rule extraction process, because all generated rules have high score.

Figure 2 shows the iris dataset with rules r_1, r_2 and r_5.

6.3 Letter Recognition

This dataset was created as a classification problem with 26 classes, one class for each letter in the English alphabet. The charactersÒ were obtained from 20 different fonts and randomly distorted to produce 20,000 unique samples presented as 16 dimensional numerical vectors. Letter X was selected as the anomaly class. SRF produced more than 15,000 rules, which were reduced by aggregation to 1955. Aggregated rules with support higher than 0.10 are presented in Table 3. The ranking of those rules is plotted at Figure 3. Its evident that the ranking given by lift and confidence boost differs substantially.

It is nearly impossible to evaluate all rules. Therefore, we have selected only those with confidence boost higher than one (202 rules) and those with lift higher than one

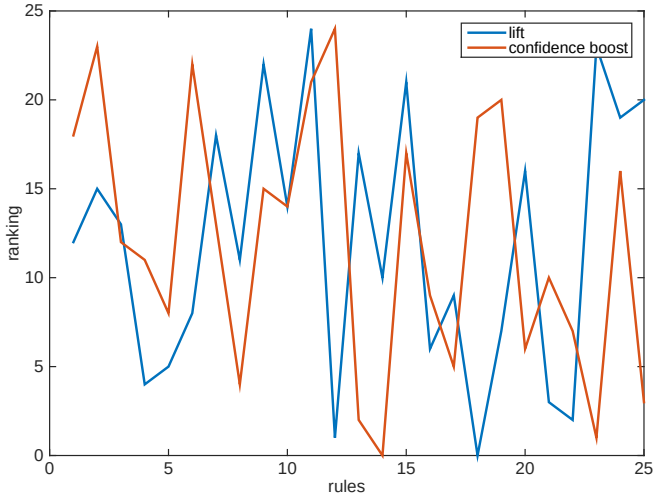

Figure 3: Ranking (higher number means better ranking) of rules from Table 3 by its lift and confidence boost.

rule	supp	lift	β
$x_2 > 14$	0.29	1.44	0.29
$x_{11} < 0.62$	0.24	0.27	1.82
$x_9 > 8.8 \wedge x_10 < 5.2$	0.21	1.67	0.96
$x_3 > 10$	0.18	0.89	0.35
$x_1 < 0.38$	0.17	0.19	0.14
$x_{13} > 6.2 \wedge x_{15} > 8.6$	0.16	0.21	0.17
$x_6 > 9.8 \wedge x_8 < 1.2$	0.15	2.39	1.91
$x_1 > 9.5 \wedge x_2 > 14$	0.15	1.10	0.22
$x_2 > 13 \wedge x_{10} > 12$	0.15	0.44	0.04
$x_4 > 7.8 \wedge x_8 < 1.2 \wedge x_9 > 7.2$	0.15	5.75	0.71
$x_8 < 1.2 \wedge x_{16} < 5.2$	0.14	1.62	0.39
$x_2 > 11 \wedge x_4 > 7.8 \wedge x_9 > 6.4$	0.13	5.32	0.88
$x_8 < 1.2 \wedge x_9 > 8.8$	0.13	8.06	1.80
$x_9 > 8 \wedge x_{10} < 5.2 \wedge x_{15} > 6.8$	0.13	2.20	1.31
$x_1 > 9.5 \wedge x_3 > 9$	0.12	1.04	0.06
$x_9 > 8.8 \wedge x_{12} < 5.5$	0.12	0.78	0.22
$x_3 > 6.1 \wedge x_{14} < 6.2 \wedge x_{15} > 7.2$	0.12	2.91	1.09
$x_1 > 4.2 \wedge x_8 < 1.2 \wedge x_{12} < 6.8$	0.12	0.55	0.05
$x_1 > 8.5 \wedge x_{10} > 12$	0.12	0.84	0.10
$x_1 > 5.5 \wedge x_7 > 7.8 \wedge x_8 < 1.2$	0.11	3.39	0.22
$x_2 < 1.5 \wedge x_9 > 7.5 \wedge x_{15} > 5.5$	0.11	3.43	0.84
$x_1 > 7.8 \wedge x_6 > 8.8 \wedge x_7 > 6$	0.11	0.29	0.02
$x_6 > 8.2 \wedge x_{15} > 7.9 \wedge x_{16} < 6.2$	0.11	0.29	0.24
$x_2 > 11 \wedge x_3 > 7.6 \wedge x_{10} < 5.2$	0.11	1.21	0.05
$x_3 > 10 \wedge x_5 > 6.2$	0.11	0.61	0.24

Table 3: Rules extracted from Letter recognition by SRF with support higher than 0.10 with their quality measures sorted by their respective supports.

(446 rules). The confidence boost selected the smaller rule set where almost all rules looked plausible. On the other hand, they missed some really interesting ones most highly rated by lift. The confidence boost tend to choose shorter more similar rules, whereas lift prefer richer and more heterogenous rules. Therefore, from our point of view the

top 10 lift rules	top 10 β rules
$x_8 < 1.8 \wedge x_{15} > 7.8$	$x_4 < 1.2 \wedge x_9 > 8$
$x_2 < 1.8 \wedge x_3 > 4.5 \wedge x_9 > 8.8$	$x_2 < 1.5 \wedge x_9 > 8.8$
$x_5 > 6.8 \wedge x_8 < 1.8 \wedge x_{15} > 7$	$x_8 < 1.2 \wedge x_9 > 8.8$
$x_5 > 5 \wedge x_9 > 8.2 \wedge x_{10} < 5.2$	$x_9 > 8 \wedge x_{14} < 5.2$
$x_4 < 2.2 \wedge x_6 > 8.2 \wedge x_9 > 7.8$	$x_{11} > 12 \wedge x_{16} < 4.5$
$x_2 < 5.6 \wedge x_3 > 7.2 \wedge x_8 < 1.2$	$x_6 > 9.8 \wedge x_8 < 1.2$
$x_1 > 4.5 \wedge x_8 < 1.8 \wedge x_{15} > 7.8$	$x_4 > 8.8 \wedge x_{14} < 5.2$
$x_7 < 5.8 \wedge x_8 < 2.5 \wedge x_{15} > 7.8$	$x_5 > 8.5 \wedge x_{14} < 5.2$
$x_2 < 4.8 \wedge x_9 > 8 \wedge x_{11} > 9.8$	$x_8 < 1.2 \wedge x_{16} > 9.9$
$x_{11} > 9 \wedge x_{14} > 13$	$x_{12} > 10 \wedge x_{16} < 5.2$

Table 4: Comparison of top 10 rules extracted from the Letter recognition dataset by SRF selected by lift and confidence boost.

best selection strategy is choosing top k rules according to the lift ranking. The top 10 rules chosen from the whole set by lift and confidence boost, regardless their support, are in Table 4.

Still there are too much rules to make some conclusions, in our future work we are going to investigate more measures of interestingness and novelty, which will hopefully help us to reduce the amount of extracted rules even more.

7 Conclusion

In this paper, we presented a novel approach for the explanation of an output of an arbitrary anomaly detector using sapling random forests. The explanation is given as conjunctions of atomic conditions, which can be viewed as antecedents of association rules. Due to an extraction method, the individual rules are short and comprehensible. The main drawback was that the rule sets for the bigger dataset were large and redundant. Therefore, we applied multiple quality measures to evaluate them and select those rules with desired properties. Performed experiments showed that no one of presented measures reflect our expectation. From the considered measures the lift looks the most promising. But this paper is just a work in progress and we don't view this observation as a final conclusion.

For our future work we would like to have a measure that will rate the novelty of a rule with respect to the set of previously selected rules. The first idea is to chose those rules that describe anomalies not covered by the already selected rules. The second idea is to select rules which may describe already covered anomalies but using completely different set of features. The last thing we would like to work on is finding a way of concatenating mined rules to make smaller yet precise rule sets.

Acknowledgement

The research reported in this paper has been supported by the Czech Science Foundation (GA ČR) grant 13-17187S.

References

[1] Aggarwal, C. C.: Outlier analysis. Springer, 2013

[2] Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD Record, volume 22, 207–216, ACM, 1993

[3] Balcázar, J. L.: Formal and computational properties of the confidence boost of association rules. ACM Transactions on Knowledge Discovery from Data (TKDD) **7(4)** (2013), 19

[4] Brin, S., Motwani, R., Ullman, J. D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, 255–264, Tucson, Arizona, USA, May 1997

[5] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Computing Surveys (CSUR) **41(3)** (2009), 15

[6] Christen, P., Goiser, K.: Quality and complexity measures for data linkage and deduplication. In: Quality Measures in Data Mining, 127–151, Springer, 2007

[7] Dang, X. -H., Micenková, B., Assent, I., Ng, R. T.: Local outlier detection with interpretation. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2013), 2013

[8] de Vries, T., Chawla, S., Houle, M. E.: Finding local anomalies in very high dimensional space. In: IEEE 10th International Conference on Data Mining (ICDM 2010), 2010

[9] Fujimaki, R., Yairi, T., Machida, K.: An approach to spacecraft anomaly detection problem using kernel feature space. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005

[10] Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., Vázquez, E.: Anomaly-based network intrusion detection: techniques, systems and challenges. Computers & Security, 2009

[11] Hawkins, D. M.: Identification of outliers, volume 11. Springer, 1980

[12] Jalali-Heravi, M., Zaïane, O. R.: A study on interestingness measures for associative classifiers. In: Proceedings of the 2010 ACM Symposium on Applied Computing, 1039–1046, ACM, 2010

[13] Karr, D.: Big data brings marketing big numbers. `[https://www.marketingtechblog.com/ibm-big-data-marketing/]`, 2012 [Online; accessed 19-June-2015].

[14] Knorr, E. M., Ng, R. T.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the International Conference on Very Large Data Bases, 1998

[15] Knorr, E. M., Ng, R. T.: Finding intensional knowledge of distance-based outliers. In: VLDB, 1999

[16] Kopp, M., Pevný, T., Holeňa, M.: Interpreting and clustering outliers with sapling random forests. In: Information Technologies – Applications and Theory Workshops, Posters, and Tutorials (ITAT 2014), 2014

[17] Lichman, M.: UCI machine learning repository. [http://archive.ics.uci.edu/ml/]. University of California, Irvine, School of Information and Computer Sciences, 2013

[18] Liu, B., Hsu, W., Chen, S.: Using general impressions to analyze discovered classification rules. In: KDD, 31–36, 1997

[19] Liu, F. T., Ting, K. M., Zhou, Z. -H.: Isolation forest. In: Eighth IEEE International Conference on Data Mining (ICDM 2008), 2008

[20] Micenková, B., Ng, R. T., Dang, X. -H., Assent, I.: Explaining outliers by subspace separability. In: IEEE 13th International Conference on Data Mining (ICDM 2013), 2013.

[21] Pevný, T., Kopp, M.: Explaining anomalies with sapling random forests. In: Information Technologies – Applications and Theory Workshops, Posters, and Tutorials (ITAT 2014), 2014

[22] Pradnya, K., Khanuja, H. K.: Article: A survey on outlier detection in financial transactions. International Journal of Computer Applications **108(17)** (December 2014), 23–25

[23] Rousseeuw, P. J., Leroy, A. M.: Robust regression and outlier detection. John Wiley & Sons, 2005

[24] Tibshirani, R., Hastie, T.: Outlier sums for differential gene expression analysis. Biostatistics, 2007