Distributed conformal anomaly detection

Ilia Nouretdinov,
Royal Holloway University of London,
i.r.nouretdinov@rhul.ac.uk

***Abstract*—Conformal approach to anomaly detection was recently developed as a reliable framework of classifying examples into normal and abnormal groups based on a training data set containing only normal examples. Its validity property is that a normal example, generated by the same distribution as the examples from the training set, is classified as anomaly with probability bounded from above by a pre-selected significance level. Parallel processing of big data may require a split of the training set into several sources. We also assume that the collection of data for two or more sources might be done in parallel and the data distribution may differ for these sources. The contribution of this work to conformal anomaly detection is studying the ways of keeping conformal validity when the training set is obtained from heterogeneous (differently distributed) sources.**

Keywords: conformal prediction, anomaly detection, distributed computing, validity.

I. Introduction

This work studies the ways of combining two things: first, reliability of machine learning predictions provided by conformal prediction framework; second, computational efficiency given by parallel programming on big data sources.

Conformal prediction framework is a way of reliable machine learning with guarantees of validity of its output in weak (i.i.d.) assumption about the data. It can be used in both supervised and unsupervised machine learning. In this task we concentrate on unsupervised task of anomaly detection.

The theory of conformal prediction for supervised learning was developed in [1] and other works. In particular, the criteria of efficiency measuring the quality of predictions were discussed in [7]. Conformal anomaly detection was introduced for wide usage in [2] and other works summarised in [3]. The work [4] supplied proper efficiency criteria. They were also applied in [5] for a multi-level version of anomaly detection. In [6] this kind of framework was also applied for clustering, including the discussion of relevant efficiency criteria.

This particular work is motivated by possible big data challenges to conformal predictors. The core detail of conformal prediction is so called conformity measure that is a kind of information similarity of an object (feature vector) and the data set of the objects of the same nature. We model the situation when the data set is so big that this conformity measure can not be calculated directly, without splitting the data set into parts. Furthermore, we assume that different parts of the data set might be not even collected and stored together. In our understanding the data may be collected in different places ('sources'), and the sources are in general case heterogeneous. There may be a significant difference in the distribution of the data from different sources.

It can be said that each of the sources is 'specialised' on some partial truth, and only being taken together, they provide full information about the true data distribution. At the same time, the testing examples to which the algorithms have to be applied, are generated by the full (mixed) distribution. It is not known to which source each of them corresponds in generating mechanism, but it is possible to compare it to data from each of the sources. Extending conformal prediction for this task is not obvious because of the risks to lose the validity and to make the predictions not very informative. We suggest a way to avoid them.

There are many areas where anomaly detection methods is applied. Problems with big data usually appear in such areas as medicine, information and other security where tracing the behaviour of a user or of a system may lead to conclusions about some unexpected effect. As a model example, we use gas consumption data provided by Energy Demand Research Project [10]. A data instance here is based on a behaviour record of an individual household consumption during a large observation time. We assume that this example is typical enough to have analogies in other areas such as catching abnormalities during usage of Internet gf Things.

The plan of the paper is following. In the background Section II we remind the basic notions of conformal prediction and conformal anomaly detection, with concentration on its validity properties. In Section III we state the challenge for validity caused by big size of the data, and suggest a way of solving it. Section IV includes some experiments. In Appendix we present some theoretical justification of the suggested algorithm.

II. Machine learning background

A. Conformal prediction

The task of machine learning is to predict a label for a new (or a testing) example x_{l+1} from a given training set of feature vectors $x_1, x_2, \ldots, x_l \in X$ supplied with labels $y_1, y_2, \ldots, y_l \in Y$.

The conformal prediction technique introduced in [1] and had many applications and extensions later. It allows to make a valid confident prediction. In conformal prediction approach for supervised learning a (feature vector, label) pair (x_i, y_i) is understood as a whole object z_i. In anomaly detection case there are no labels and $z_i = x_i$.

The core detail of conformal predictor in both cases for a *conformity measure (CM) A* that is a measure information similarity of an object z, which is usually a feature vector, and a set U of objects of the same nature. In other way it can be said that CM estimates relative typicalness of the objects $z_1, \ldots, z_{l+1}$ with respect to each other:

$$\alpha_i = A\left(z_i, \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{l+1}\}\right).$$

Some previous works such as [1], [4], [7] also use the term *non-conformity measure (NCM)* which differs from conformity measure in the sign.

In conformal anomaly detection the conformal predictor assigns p-values (or the values of a *test for randomness*)

$$p = p(z_1, \ldots, z_{l+1}) = \frac{card\{i = 1, \ldots, l+1 : \alpha_i \leq \alpha_{l+1}\}}{l+1} \quad (1)$$

that measures how likely a new example z_{l+1} is to be generated by the same distribution as the previous examples $z_1, \ldots, z_l$.

B. Validity properties

The validity property states that if the sequence of examples $z_1, \ldots, z_{l+1}$ are really generated by an i.i.d. (power) distribution then for any significance level ε, the probability that $p < \varepsilon$ is at most ε.

In unsupervised case, the prediction set $R^\varepsilon \subset X$ is the set of z s.t.

$$p(z) = p(z_1, \ldots, z_l, z) > \varepsilon.$$

Validity implies that this R^ε covers the real example z_{l+1} with probability at least $1 - \varepsilon$. When the example z_{l+1} becomes known, it is reported as an anomaly if $p(z_{l+1}) < \varepsilon$ because the probability of this event is below the selected significance level ε.

As follows from the validity, if z_{l+1} is really generated by the same stochastic mechanism as $z_1, \ldots, z_l$, then it is (wrongly) classified as anomaly with probability at most ε. At the same time we wish an example z_{l+1} breaking i.i.d. assumption to be classified as anomaly as often as possible. Therefore the smaller is the prediction set the more efficient is the prediction.

C. Deviations from validity

If the formula 1 of p-value calculation is modified or generalised in some way, the challenge is to keep its validity property, otherwise the outputs of conformal prediction would not be reliable. Ideally, probability that $p < \varepsilon$ should be very close to ε. If this probability is essentially smaller than ε this means the results are reliable but too conservative. Conservativeness is an indirect but important indicator of prediction sets being unnecessarily large. Indeed, if the allowed level ε of normal examples classified as anomalies is not reached, then it is very possible that some true anomalies are also left undetected. On the other hand, if probability of this event is significantly larger than ε, this means that there is no reliability at all, which is much worse than the loss of some efficiency.

Therefore, the general question is to check how much the empirical distribution of p deviates from the uniform distribution on [0,1]. An on-line vesion of such testing was earlier presented in [8] but here we are testing algorithms, not the data sets themselves. Also two kinds of deviations are principally different for us now. Therefore for simplicity we will process the data in batch (not on-line) mode and measure just the average p-values.

For a given significance level ε three versions are possible:

1) Exact validity: $Prob\{p \leq \varepsilon\} = \varepsilon$;
2) Conservative validity: $Prob\{p \leq \varepsilon\} < \varepsilon$;
3) Invalidity: $Prob\{p \leq \varepsilon\} > \varepsilon$.

Invalidity means that p-value is wrong as a measure of significance. Conservative validity is allowed, but it reflects that a test for randomness not very sensitive.

What we can do in practice is to look at the empirical distribution of p-values assigned to testing examples. Ideally (in the case of exact validity) it should be uniform, as it is for smoothed version of non-parallelised conformal prediction [1]. However in this work we do not do special smoothing. For the big data size there is usually no practical difference between smoothed and non-smoothed versions, so it is convenient to get rid of this non-deterministic element.

In order to get an aggregated ε-independent p-value we measure averaged p-value (APV) suggested in [4] in a slightly different context. If p-values are distributed uniformly, its value has to be insignificantly different from $\frac{1}{2}$.

This criteria may be considered as paying equal attention to small and large ε, while in practice only its small values are usually important: significance levels typically used in statistics are no larger than 0.05, and it is desirable to consider even smaller ones such as 10^{-2} or 10^{-4}. Therefore we also consider a modified criterion that is average logarithm of p-value (ALPV), with the expected value of -1 for uniform distribution. This criterion gives larger weight to small signifance levels. For example, if the p-values tend to concentrate around $\frac{1}{2}$ this is not reflected in APV but considered conservative by ALPV.

III. Aggregating data sources

A. Big data challenges

Some algorithms of machine learning do have special, and usually approximate, modifications applicable to big data. An example is Cascade Support Vector Machines [9] created for the case when the number of data examples (feature vectors) is too large. The principal challenge is that underlying SVM is using a matrix of inter-example similarities, so the load of memory is proportional to the square of the number of training examples, which puts an essential limit on the number of examples which may be processed together.

A specific challenge related to conformal predictor is related to its core detail, conformity measure function A. One of its two arguments is a whole set, this may cause problems if the training set becomes very big. We can just assume that an upper limit on the set size for this function is determined on the power of one processor, and on the complexity of this conformity measure. In the example which we will consider, CM is based on k-Nearest-Neighbour algorithm, which also requires a storage of the distance matrix.

Assuming that either the data or a core data-based matrix is out of memory, this step requires parallelization by splitting of this set into several subsets. However as we will see further, doing this in a straightforward way may lead to some loss of reliability which was guaranteed by validity properties.

We also assume another possible cause of parallelization: data collection was done independently by several groups, and it even may be never stored together. Also it is possible in

practice that each of the collecting group may have its own specialization collecting data of a concrete subtype, or related to the collection place. This makes the problem harder. First, in general case it is impossible to make a completely random split. Second, if there is a difference in the distribution of different data source, it would be unfair to restrict us just to one of these sources, even as an approximation.

In the following we will assume that the training set $U = \{z_1, \ldots, z_l\}$ is split into two parts of equal size called U_1 and U_2, and this in general case is not always a random split. Our basic assumption is the following. The true data distribution P is a mixture of P_1 and P_2,

$$P = \frac{P_1 + P_2}{2}.$$

The data source U_1 is randomly generated by P_1, and the source U_2 by P_2. However, for a testing example z_{l+1} we do not know which of these sources it is coming from, so it is generated by the mixed distribution P. The task is to estimate its randomness with respect to the union of U_1 and U_2, in the restriction that simultaneous access to the sources U_1 and U_2 is impossible.

The more general case of several sources of non-equal size is defined by analogy and can be found in Appendix.

B. Averaging tests for randomness (AT)

Let us start with a straightforward way of aggregation is to calculate p_1 and p_2 are calculated for the same testing example but for different training sets, and to try aggregating them. This means splitting Equation 1 into two:

$$p_1 = \frac{card\left\{i = 1, \ldots, \frac{l}{2}, l+1 : A(z_i, U_1) \le A(z_{l+1}, U_1)\right\}}{\frac{l}{2}+1} \quad (2)$$

$$p_2 = \frac{card\left\{i = \frac{l}{2}+1, \ldots, l, l+1 : A(z_i, U_2) \le A(z_{l+1}, U_2)\right\}}{\frac{l}{2}+1} \quad (3)$$

Can these p_1 and p_2 be combined into some approximation of p given by Equation 1? The average test value

$$p^{AT} = \frac{p_1 + p_2}{2}$$

can be considered as an approximation to some extent: the same conformity scores are summarised onto the p-value, but each of these conformity score is approximately calculated on the base of smaller number of examples than it should be.

This method has no guarantees of validity. If the split $U_1 \cup U_2$ is random (homogeneous), it is typically valid but tends to be conservative for small values of ε. However, if the split is heterogeneous, then the problem of invalidity can appear as well. This is evident on a 'completely heterogeneous' example of the sources. Indeed, assume that P_1 and P_2 distributions having non-overlapping support sets that are so disjointed that each example generated by P_1 has the minimal possible p-value $\frac{1}{l+1}$ with respect to the training set generated by P_2 and vice versa. This will lead to averaged p-value of a new P-generated example being approximately unoformly distributed on $\left(0, \frac{1}{2}\right)$ instead of $(0,1)$. In particular, it is limited from above by

$$\frac{1 + \frac{1}{l+1}}{2} = \frac{1}{2} + \frac{1}{2l+2}$$

which definitely contradicts the validity for any $\varepsilon > \frac{1}{2} + \frac{1}{2l+2}$.

C. Maximizing tests for randomness (MT)

Another straightforward way maximizing the testing values:

$$p^{MT} = \max\{p_1, p_2\}.$$

This is an easy straightforward way to achieve validity. It also has a natural meaning: a new example z is typical with respect to the union of two sources if is typical with respect to at least one of them. Usually p-values produced this way are very conservative, unless P_1 and P_2 are completely disjointed as in the example considered above.

D. Maximizing conformities (MC)

The suggested solution is to move the step of maximization back. Instead of maximizing the p-values themselves, let us maximize the estimates of new example's conformity. The aggregated p-value is defined as:

$$p^{MC} = \tilde{p} = (\tilde{p}_1 + \tilde{p}_2)/2$$

$$\tilde{p}_1 = \frac{card\left\{i = 1, \ldots, \frac{l}{2}+1 : A(z_i, U_1\}) \le A(z, U_1, U_2)\right\}}{\frac{l}{2}+1} \quad (4)$$

$$\tilde{p}_2 = \frac{card\left\{i = 1, \ldots, \frac{l}{2}+1 : A(z_i, U_2\}) \le A(z, U_1, U_2)\right\}}{\frac{l}{2}+1} \quad (5)$$

where

$$A(z, U_1, U_2) = \max\{A(z, U_1), A(z, U_2)\}.$$

A theoretical justification of this formula is given in Appendix where it is shown to keep validity at least in some asymptotic sense, if the conformity measure has a natural sense of density approximation, up to a monotonic change. There it is also presented in a more general form, covering the case on several sources of non-equal size.

These methods require to store more information than just p-values, but it is still applicable to split data sets, because U_1 and U_2 are used separately of each other, and only conformity scores have to be stored. Furthermore, it is not necessary to keep the conformity score assigned to training examples in the original order, it is quite enough to have the overall empirical distribution of them.

E. An example

An example of three ways compared to each other is presented in Fig. 1.

We assume that the size of two sources U_1 and U_2 is 9, and there is one testing example z. It is replicated twice on the picture because it may get different conformity values when compared to two data sources. In this example they are

$$p_1 = \frac{4+1}{9+1} = 0.5; p_2 = \frac{1+1}{9+1} = 0.2.$$

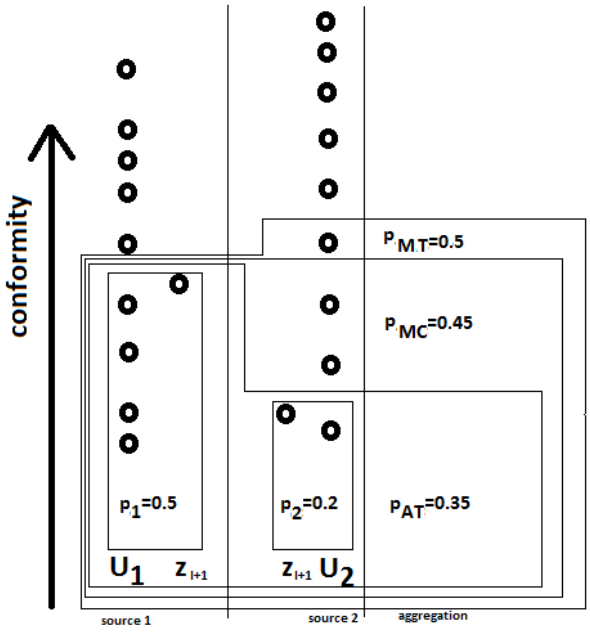

Fig. 1. Original and aggregated p-values as explained in Sec. III-E

Feature No.	Feature name
1	Inclusive number of days between first and last advances
2	Lowest advance value
3	Highest advance value
4	Average advance value
5	Number of advances expected based on feature 1
6	Number of available advances
7	Ratio of features 5 and 6
8	Time of use identifier

TABLE I
METADATA FEATURES

INPUT: proper training set $w_1,\ldots,w_h$
INPUT: calibratlion set $z_1,\ldots,z_l$
INPUT: testing example z_{l+1}
INPUT: number k of neighbours
FOR $i=1,\ldots,l+1$
FOR $j=1,\ldots,h$
$d_{ij}=|z_i-w_j|^2$
END FOR
let $d'_{i1},\ldots,d'_{ih}$ be $d_{i1},\ldots,d_{ih}$ sorted in ascending order
$\alpha_i=\frac{1}{d'_{i1}+\ldots+d'_{ik}}$
END FOR
OUTPUT $p(z_{l+1})=\frac{card\{i=1,\ldots,l+1:\alpha_i\leq\alpha_{l+1}\}}{l+1}$

TABLE II
SCHEME OF THE INDUCTIVE CONFORMAL PREDICTOR APPLIED IN THE EXPERIMENTS.

Naturally, the averaged value $p^{AT}=0.35$ and the maximal one is $p^{MT}=0.5$. On the picture we draw for each of them a boundary, including the example considered 'stranger or equal' than the new one after the aggregation. The same is done for

$$p^{MT}=\frac{(4+1)+(3+1)}{2\times(9+1)}=0.45$$

where the example were selected according to the highest of two thresholds, earlier used for calculating p_1 and p_2.

IV. EXPERIMENTS

A. Data

For the experiments we use gas consumption data taken from Energy Demand Research Project [10]. The data structure is following.

- **geography data:** various characteristics of 16,249 households (14 attributes)
- **metadata:** summary of the energy usage for these households (11 attributes for electricity and 11 for gas if it is used).
- **edrp-elec** and **edrp-gas** big files showing energy consumption for households at various types.

For our aims we use 8 numerical metadata features for 8,703 gas users, listed in Table I.

These features are put into logarithmic scale and normalised. In some experiments we also use ACORN category from geography data, showing the type of household by social characteristics of its owners or tenants. Each example is assigned one of 6 categories.

B. General settings

As discussed in the Appendix, a suitable conformity measure should approximate the local density of data distribution (up to a monotonic transformation). We use k-nearest-neighbours method as the simplest straightforward method of approximation density. Conformity of an example is calculated as the inverse value of the average distance to k nearest neighbours from the remaining data set where k plays role of a 'focusing' parameter. The intuition of this approximation is: the more dense is the data concentration in the area around a specific example, the smaller is its average distance to nearest neighbours. Another simple approach can be based on approximation conformity by density function, the comparison in [4] in the context of conformal anomaly detection, shows similarity of its best results to best k-nearest-neighbours results in typical cases.

We process data in so called *inductive* mode presented in [1]. It means that a special part of data called *proper training set* is left aside the learning and is used only in calculation conformities of the remaining *calibration* examples. This makes computation time practically same as one needed for calculation of the distance matrix that is the most computationally expensive step.

For convenience, all the steps are presented together in the Table II. This algorithm is applied with the number of neighbours: $k=100$.

C. Distributed computing

In our experiments we split the data set into two sources of equal size. Each of them may have its own distribution, however we assume within one source the order of examples is random. The sizes of proper training and calibration sets within each source are equal and the same as the number of testing examples. However, testing examples from both sources are mixed together into a joint testing set. This can be

Average p-value	Average log. of p-value	Interpretation
$< \frac{1}{2}$	< -1	invalid results
$= \frac{1}{2}$	$= -1$	valid and efficient results
$> \frac{1}{2}$	> -1	valid conservative results

TABLE III
UNDERSTANDING THE RESULTS OF EVALUATION

summarised as following. Let h be the set size, P_1 and P_2 the distributions within two sources. There are:

- proper training set 1 of size h generated by P_1^h;
- calibration set 1 of size h generated by P_1^h;
- proper training set 2 of size h generated by P_2^h;
- calibration set 2 of size h generated by P_2^h;
- $2h$ testing examples, generated by

$$P^{2h} = \left(\frac{P_1 + P_2}{2}\right)^{2h}.$$

For each of the testing examples, Inductive Confidence Machine (as presented in Table II) has to be run twice. Each of 'new' (testing) examples is compared in parallel to the first and to the second source. Three comparable approaches of aggregation were described in Sections III-B–III-D:

1) averaging tests (AT);
2) maximizing tests (MT);
3) maximizing conformities (MC).

Remind that in the ways 1–2 we need to collect only p-values. For the last way we also have to keep in the memory conformity scores for all the examples.

D. Measuring deviations from validity

For measuring deviations we use a joint ε-free interpretation of criteria earlier presented in Section II-C. It is summarised in the Table III. Validity in sense of this table does not imply validity for any ε. It just means that invalidity is invisible with this way of its measurement. Remind also that keeping validity is strictly the first priority, while conservativeness has to be reduced where possible.

E. Results

The results are presented in Table IV. Source size means the number of examples randomly taken from each source for proper training set. The same number is taken for calibration and testing sets. For example, if source size is 1000, then the first ICP learns on 1000 proper training and 1000 calibration examples form the first source, the second ICP learns on 1000 proper training and 1000 calibration examples form the second source, they both assign p-value (p_1 and p_2 respectively) to each of 2000 testing examples (taken from both sources), then these p-values are aggregated to p by one of three rules (AT, MT, MC) defined earlier.

We consider three types of splitting the data:

1) Random (homogeneous) split.
2) Split by ACORN category of household (valued $1, \ldots, 6$).
3) Split by a feature, using its median value as a threshold.

Size h	Source split	AT APV	MT APV	MC APV	AT ALPV	MT ALPV	MC ALPV
1000	random	.500	.513	.512	-0.993	-0.958	-0.971
500	cat.1/3	.495	.524	.524	-0.991	-0.926	-0.935
1000	cat.1,5/3,4	.496	.520	.519	-1.001	-0.947	-0.961
1000	cat.1,4/3,5	.492	.525	.523	-1.005	-0.936	-0.950
1000	low/high f1	.363	.528	.510	-1.244	-0.840	-0.962
1000	low/high f3	.354	.512	.501	-1.276	-0.902	-0.991
1000	low/high f4	.411	.510	.502	-1.158	-0.924	-0.990
1000	low/high f5	.363	.528	.510	-1.244	-0.840	-0.962
1000	low/high f6	.360	.535	.509	-1.243	-0.815	-0.962
1000	low/high f7	.436	.520	.532	-1.109	-0.922	-0.920

TABLE IV
RESULTS

The first way of splitting contains no heterogeneity. The second way is expected to contain it as far as different categories show different behaviour. In the third way the splitting procedure naturally leads to very high level of heterogeneity. All the results are averaged over 50 random reshuffles.

As we can see, in case of a purely homogeneous (random) split the best way of aggregating p-value is just averaging them. However, it still shows slight conservativeness concentrated at small values of ε (AT-ALPV is -0.993 instead of -1). Avoiding this may be the topic of a future study, concentrated on homogeneous case. But there is no tendency of breaking the validity. In all the rest experiments we see it: all AT-APVs are smaller than $\frac{1}{2}$ and all but one AT-ALPVs are smaller than -1. Depending on the way of splitting, this tendency may be smaller or larger. In all the experiments MT-APV and MT-ALPV show conservativeness shown by another sort of deviation from $\frac{1}{2}$ and -1 respectively. In all the experiments but the last one it is reduced by changing MT to MC. It is also interesting that the improvement is relatively small in APV but very essential in ALPV. As well, the fault of the last experiment is very small in ALPV (-0.922 is changed to -0.920).

V. CONCLUSION AND DISCUSSION

In this work we suggest a MC way of aggregation of p-values for conformal prediction on big and heterogeneously splitted data. As it is shown in the Appendix, this can be used as a basic method of aggregating p-values for conformal prediction from two or more sources. We have concentrated on the case of two sources of equal size. This made experimental comparison with other approaches more transparent. The experiments have shown that averaging p-values is dangerous for the validity if the data sources are heterogeneous, while maximizing p-value is very conservative. The method we suggested is a compromise: the validity property is kept, while the conservativeness is essentially smaller than for aggregating p-values by maximizing them. However the general case presented in Proposition 1 (see Appendix) is applicable as well to a more general case (several sources, different size) and extendable to supervised learning.

Its experimental validation on really big data sets is a topic for the future work. What we would like to concentrate in the future is growing area of internet of things such as medical sensors. It is important to detect anomalous state of a patient

and/or the devices shown by the measurements provided by sensors, to produce an alarm. The practice shows that the alert criteria have to be based on the analysis of essential data collections from various sources, otherwise they may be too approximate in a specific patient's situation.

VI. Acknowledgments

This work was supported by Technology Integrated Health Management (TIHM) project awarded to the School of Mathematics and Information Security at Royal Holloway as part of an initiative by NHS England supported by InnovateUK.

It was also supported by European Union grant 671555 ("ExCAPE"), EPSRC grant EP/K033344/1 ("Mining the Network Behaviour of Bots"), Thales grant ("Development of automated methods for detection of anomalous behaviour"), and by the National Natural Science Foundation of China (No.61128003) grant. We are grateful to Lars Carlsson for the initial motivation of the topics and useful discussions, and to Energy Demand Research Project for providing the data [10].

References

[1] Vovk, V., Gammerman, A., Shafer, G. Algorithmic Learning in a Random World. Springer, 2005

[2] Laxhammar, R., Falkman, G. Sequential conformal anomaly detection in trajectories based on hausdorff distance. Proceedings of the 14th International Conference onInformation Fusion (FUSION), 2011, pp.1–8.

[3] Laxhammar, R. Conformal Anomaly Detection. Detecting abnormal trajectories in surveillance applications. Doctoral dissertation, 2014. University of Skövde 2014, Sweden. https://www.diva-portal.org/smash/get/diva2:690997/FULLTEXT02.pdf

[4] Smith, J., Nouretdinov, I., Craddock, R., Offer, C., Gammerman, A. Anomaly Detection of Trajectories with Kernel Density Estimation by Conformal Prediction. Artificial Intelligence Applications and Innovations: AIAI2014 Workshops. Rhodes, Greece: Springer, pp. 271–280.

[5] Smith, J., Nouretdinov, I., Craddock, R., Offer, C., Gammerman, A. Conformal Anomaly Detection of Trajectories with a Multi-class Hierarchy. 2015 Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings. Vol. 9047, p. 281-290 10 p.

[6] Cherubin, G., Nouretdinov, I., Gammerman, A., Jordaney, R., Wang, Z., Papini, D., Cavallaro, L. Conformal Clustering and Its Application to Botnet Traffic. Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings. Vol. 9047, p. 313-322 10 p

[7] Vovk, V., Fedorova, V., Nouretdinov, I., Gammerman, A. Criteria of efficiency for conformal prediction. 2014 On-line COMPRESSION Modelling Project (New Series), 19 p.

[8] Fedorova, V., Gammerman, A., Nouretdinov, I., Vovk, V. Plug-in martingales for testing exchangeability on-line. Proceedings of the 29th International Conference on Machine Learning (ICML-2012), Omnipress, pp. 1639–1646.

[9] Graf, H., Cosatto, E., Bottou, L., Durdanovic, I., Vapnik, V. Parallel support vector machines: The cascade SVM. In: NIPS (2004).

[10] AECOM Building Engineering, Energy Demand Research Project: Early Smart Meter Trials, 2007-2010 [computer file]. Colchester, Essex: UK Data Archive [distributor], November 2014. SN: 7591. Energy Demand Research Project: Early Smart Meter Trials, 2007-2010, UKDA study number:7591. Principal Investigator: AECOM Building Engineering Data Collector: Centre for Sustainable Energy. Sponsor: Department of Energy and Climate Change. Distributed by UK Data Archive, University of Essex, Colchester. November 2014. http://doc.ukdataservice.ac.uk/doc/7591/mrdoc/UKDA/UKDA_Study_7591_Information.htm

Appendix: theoretical analysis and justification

Assume that data U comes from several sources $U_1,\dots,U_k$ with weights $w_1,\dots,w_k$ s.t. $w_1+\dots+w_k=1$.

The data in a source U_i is generated by an i.i.d. distribution P_i^*, and together they form a 'full' distribution P^* (the star here means a power distribution) s.t.

$$P=\frac{w_1P_1+\dots+w_kP_k}{w_1+\dots+w_k}$$

and new objects are generated by P.

In this analysis we consider some asymptotical tendention, assuming that the amount of data in each of the sources is representative enough, the difference between true and empirical distribution is as low as needed. More concetely, we assume that:

- The space X is discrete (finite) but if needed large enough.
- Each bag U_i is big and representative for P_i, so we can assume that P_i and the uniform distribution on U_i practically coincides.
- As well, the change of this distribution by adding one example may be neglected.
- A conformity measure $A(x,U)$ is the 'optimal' one i.e. it is equivalent to the local density:

$$\frac{card\{u\in U : u=x\}}{card(U)}$$

This allows to consider the conformity score $A(x,U_i)$ as a practical equivalent of the density function $P_i(x)$. Note that a non-conformity measures are equivalent if they are reducible to each other by monotonic transformation, therefore such methods as Nearest Neighbours are suitable for this if X is a vector space.

Proposition 1. *In the assumptions of this Appendix, a valid test for randomness is*

$$\tilde{p}(z)=\sum_{j=1}^{k} w_j\tilde{p}_j(z)$$

where

$$\tilde{p}_j(z)=P_j\left\{x : w_jP_j(x)\le\max_{i=1}^{k}\{w_iP_i(z)\}\right\}.$$

Proof:

Assume that an example x is generated in two steps: first, $i(x)\in\{1,\dots,k\}$ is generated according to the distribution $W=(w_1,\dots,w_k)$, then x itself is generated by $P_{i(x)}$. Set the conformity measure to $A((i,x),U)=w_iP_i(x)$ (using our assumption that P_i is recoverable from U_i with required precision). In this case the corresponding conformal p-value would be:

$$p((i,z))=\sum_{j=1}^{k} w_jP_j\left\{x : w_jP_j(x)\le w_iP_i(z)\right\}$$

$$\le\sum_{j=1}^{k} w_jP_j\left\{x : w_jP_j(x)\le\max_{i=1}^{k}\{w_iP_i(z)\}\right\}=\tilde{p}(z)$$

The last estimate does not depend on i and can be used as a p-value for z.