Change Detection Based on the Distribution of p-Values

Katharina Tschumitschew[1] and Frank Klawonn[1,2]

Abstract Non-stationarity is an important aspect of data stream mining. Change detection and on-line adaptation of statistical estimators is required for non-stationary data streams. Statistical hypothesis tests may also be used for change detection. The advantage of using statistical tests compared to heuristic adaptation strategies is that we can distinguish between fluctuations due to the randomness inherent in the underlying distribution while it remains stationary and real changes of the distribution from which we sample. However, the problem of multiple testing should be taken into account when a test is carried out more than once. Even if the underlying distribution does not change over time, any test will erroneously reject the null hypothesis of no change in the long run if we only carry out the test often enough. In this work, we propose methods which account for the multiple testing issue and consequently improve reliability of change detection. A new method based on the information about the distribution of p-values is presented and discussed in this article as well as classical methods such as Bonferroni correction and the Bonferroni-Holm method.

1 Introduction

One of the most important aspects in data stream analysis is that in most applications the underlying data generating process does not remain static, i.e. the underlying probabilistic model cannot be assumed to be stationary. The changes in the data structure may occur over time. Dealing with non-

[1] Department of Computer Science, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, D-38302 Wolfenbüttel, Germany,
{katharina.tschumitschew,f.klawonn}@ostfalia.de
[2] Bioinformatics and Statistics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweig, Germany, frank.klawonn@helmholtz-hzi.de

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 191–203.
springerlink.com

stationary data requires change detection and on-line adaptation. Different kinds of non-stationarity have been classified in [1]:

- Changes in the data distribution: the change occurs in the data distribution in general. For instance, the mean or the variance of the data distribution may change over time.
- Changes in concept: here concept change refers to changes of a target variable. A target variable is a variable whose values we try to predict based on the model estimated from the data, for instance, for linear regression concept drift refers to the change of the coefficients of the linear model which is used to predict the target variable. Concept change can be further distinguished in the following way:
  - Concept drift: concept drift describes gradual changes of the concept. In statistics, this is usually called structural drift.
  - Concept shift: concept shift refers to an abrupt change which is also referred to as structural break.

In the following, we do not differentiate between concept drift and shift for two reasons. First of all, in both cases the relation between the predictor attributes and the target variable will be changed anyway. Secondly, we can only observe or sample the data at discrete time points, so that it does not matter whether we interpret the changes between two time points as a discontinuous jump in terms of concept shift or as a smooth transition between two time points which we cannot describe or observe in detail, because we have data between two discrete time points.

Real world applications for non-stationary data can be found for instance in stock market or weather prediction, change of protein structures through mutation or the buying behaviour of customers of an on-line store. Since non-stationary data models significantly affect the accuracy of prediction, the fact of concept drift should be taken into account by on-line learning. Hence the effective treatment of non-stationarity is an important problem in machine learning. Therefore change detection and on-line adaptation for data stream mining techniques are required for non-stationary data streams. Various strategies to handle non-stationarity are proposed, see for instance [6] for a detailed survey of change detection methods. Statistical hypothesis tests may also be used for change detection. Since we are working with data streams, it is required that either the calculations for the hypothesis tests can be carried out in an incremental way or time window techniques should be used. Hypothesis tests could be applied to change detection in two different ways (for detailed survey see [12]):

- Change detection through incremental computation of the tests: by this approach the test is computed in an incremental fashion. For instance, the χ^2-test and the t-test (for precise definitions see for example [10]) render themselves easily to incremental computations (on-line adaptation of these tests is described in [12]). A low p-value for the comparison of the

data distributions at different time points – in the case of the χ^2-test – or comparison of the mean values – in the case of the t-test – would indicate a change in the data stream.
- Time window techniques: by this approach the data stream is divided into time windows. A sliding window can be used as well as non-overlapping windows. In order to detect potential changes, we need either to compare data from an earlier window with data from newer one or to test only the new data (for instance, whether the data follow a known or assumed distribution).

However, the problem of multiple testing should be taken into account when more than one hypothesis is tested simultaneously. The more hypotheses are tested, the more likely the null hypothesis of no change will be erroneously rejected, even if the underlying distribution does not change over time. In this work we present different approaches to solve this problem. One way is the application methods that account for multiple testing like the well known Bonferroni correction and the Bonferroni-Holm method. Furthermore, we propose a new approach based on the information about the distribution of p-values.

This paper is organised as follows. The problem of multiple testing is explained in Section 2. Two classical methods to handle the problem of multiple testing are also described in this section. In Section 3 the theoretical background on p-values is given and a new approach based on the distribution of p-values under the null hypothesis is introduced. Examples are discussed in the experimental section 4.

2 Multiple Testing

Multiple testing refers to the application of a number of tests simultaneously. Instead of a single null hypothesis, tests for a set of null hypotheses H_0, $H_1, \ldots, H_n$ are considered. These null hypotheses do not have to exclude each other.

An example for multiple testing is a test whether m random variables $X_1, \ldots X_m$ are pairwise independent. This means the null hypotheses are $H_{1,2}, \ldots, H_{1,m}, \ldots,$ $H_{m-1,m}$ where $H_{i,j}$ states that X_i and X_j are independent. Multiple testing leads to the undesired effect of cumulating the α-error.

Definition 1. The α-error α is the probability to reject the null hypothesis erroneously, given it is true.

Choosing $\alpha = 0.05$ means that in 5% of the cases the null hypothesis would be rejected, although it is true. When k tests are applied to the same sample, then the error probability for each test is α. Under the assumption that the null hypotheses are all true and the tests are independent, the probability that at least one test will reject its null hypothesis erroneously is

$$\begin{aligned} P(\ell \geq 1) &= 1 - P(\ell = 0) \\ &= 1 - (1-\alpha) \cdot (1-\alpha) \ldots \cdot (1-\alpha) \\ &= 1 - (1-\alpha)^k. \end{aligned} \tag{1}$$

ℓ is the number of tests rejecting the null hypothesis.

A variety of approaches have been proposed to handle the problem of cumulating the α-error. In the following, two common methods will be introduced shortly.

The simplest and most conservative method is Bonferroni correction [9]. When k null hypotheses are tested simultaneously and α is the desired overall α-error for all tests together, then the corrected α-error for each single test should be chosen as $\tilde{\alpha} = \frac{\alpha}{k}$. The justification for this correction is the inequality

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i). \tag{2}$$

For Bonferroni correction, A_i is the event that the null hypothesis H_i is rejected, although it is true. In this way, the probability that one or more of the tests rejects its corresponding null hypothesis is at most α. In order to guarantee the significance level α, each single test must be carried out with the corrected level $\tilde{\alpha}$.

Bonferroni correction is a very rough and conservative approximation for the true α-error. One of its disadvantages is that the corrected significance level $\tilde{\alpha}$ becomes very low, so that it becomes almost impossible to reject any of the null hypotheses.

The simple single step Bonferroni correction has been improved by Holm [7]. The Bonferroni-Holm method is a multi-step procedure in which the necessary corrections are carried out stepwise. This method usually yields larger corrected α-values than the simple Bonferroni correction.

When k hypotheses are tested simultaneously and the overall α-error for all tests is α, for each of the tests the corresponding p-value is computed based on the sample x and the p-values are sorted in ascending order.

$$p_{[1]}(x) \leq p_{[2]}(x) \leq \ldots \leq p_{[k]}(x) \tag{3}$$

The null hypotheses H_i are ordered in the same way.

$$H_{[1]}, H_{[2]}, \ldots, H_{[k]} \tag{4}$$

In the first step $H_{[1]}$ is tested by comparing $p_{[1]}$ with $\frac{\alpha}{k}$. If $p_{[1]} > \frac{\alpha}{k}$ holds, then $H_{[1]}$ and the other null hypotheses $H_{[2]}, \ldots, H_{[k]}$ are not rejected. The method terminates in this case. However, if $p_{[1]} \leq \frac{\alpha}{k}$ holds, $H_{[1]}$ is rejected and the next null hypothesis $H_{[2]}$ is tested by comparing the p-value $p_{[2]}$ and the corrected α-value $\frac{\alpha}{k-1}$. If $p_{[2]} > \frac{\alpha}{k-1}$ holds, $H_{[2]}$ and the remaining null hypotheses $H_{[3]}, \ldots, H_{[k]}$ are not rejected. If $p_{[2]} \leq \frac{\alpha}{k-1}$ holds, $H_{[2]}$ is rejected and the procedure continues with $H_{[3]}$ in the same way.

The Bonferroni-Holm method tests the hypotheses in the order of their p-values, starting with $H_{[1]}$. The corrected α_i-values $\frac{\alpha}{k}, \frac{\alpha}{k-1}, \ldots \alpha$ are increasing. Therefore, the Bonferroni-Holm method rejects at least those hypotheses that are also rejected by simple Bonferroni correction, but in general more hypotheses can be rejected.

During change detection instead of the common significance level α, the Bonferroni correction or Bonferroni-Holm method should be used in order to avoid the multiple testing problem. However, the streaming nature of the data should be taken into account and it is therefore impossible to hold all the obtained p-values in the memory. Furthermore, the number of tests to be carried out is not known in advance. Thus, a time window technique-based approach should be used, such for instance as a sliding window or non-overlapping time windows.

3 Meta p-values

Another possibility to solve the problem of multiple testing during change detection is to study the behaviour of the obtained p-values. Several authors have analysed properties of p-values. For instance, Gibson and Pratt (see [5]) provided an interpretation and methodology for p-values, Sackrowitz and Samuel-Cahn [8] analysed the stochastic behaviour of p-values. Donahue in [4] studied the distribution of p-values under the alternative hypothesis. In [2], the authors focus on the median of the p-value under the alternative hypothesis.

Definition 2. The p-value is the probability to obtain a value of the test statistic as extreme as, or more extreme than (depending on the alternative hypothesis) the observed value of the test statistic given the null hypothesis is true.

Hence, in the case of continuous test statistics for a right tailed test the p-value is calculated as

$$p = \Pr\left(T \geq t | H_0\right) = 1 - F_T(t) \tag{5}$$

and for a left tailed test as

$$p = \Pr\left(T \leq t | H_0\right) = F_T(t) \tag{6}$$

where $F_T(t)$ is the cumulative distribution function for the test statistic T under the assumption that the null hypothesis H_0 is true.

In the case of a two tailed test, the p-value is the total area under both tails with an area of $\frac{p}{2}$ in each tail. Therefore, if the observed value falls into the one-tailed area, the area of this tail has to be doubled and the other tail can be ignored.

$$p = \begin{cases} 2 \cdot \Pr\left(T \leq t | H_0\right), \text{if } t \leq q_{0.5}^T \\ 2 \cdot \Pr\left(T \geq t | H_0\right), \text{otherwise.} \end{cases} \tag{7}$$

As the Equations (5), (6) and (7) show, the p-value is a function of a random variable and hence a random variable itself. An obvious question is: how are the p-values distributed under the null hypothesis and how under the alternative hypothesis? First, the distribution of p-values is analysed when H_0 is true (see [4, 8]).

Theorem 1. *Given the null hypothesis is true, the p-values of a continuous test statistic T follow a uniform distribution on the unit interval* $[0, 1]$.

Proof. Let p be the achieved p-value and t the calculated test statistic with $F_P(p|H_0)$ and $F_T(t)$ being the corresponding cumulative distribution functions under H_0. Also, let $F_T^{-1}(\gamma)$ be the inverse function of $F_T(t)$, so that $F_T\left(F_T^{-1}(\gamma)\right) = \gamma$ for all $\gamma \in [0, 1]$. Then, for a right tailed test the following holds

$$\begin{aligned} F_P(p|H_0) &= \Pr\left(P \leq p|H_0\right) \\ &= \Pr\left(1 - F_T\left(t\right) \leq p|H_0\right) \\ &= \Pr\left(F_T\left(t\right) \geq \left(1-p\right)|H_0\right) \\ &= 1 - \Pr\left(F_T\left(t\right) \leq \left(1-p\right)|H_0\right) \\ &= 1 - F_T\left(F_T^{-1}\left(1-p\right)\right) \\ &= 1 - (1-p) = p \end{aligned} \tag{8}$$

For a left tailed test corresponding to Equation (6) the distribution function of the p-value is as follows

$$\begin{aligned} F_P(p|H_0) &= \Pr\left(P \leq p|H_0\right) \\ &= \Pr\left(F_T\left(t\right) \leq p|H_0\right) \\ &= F_T\left(F_T^{-1}\left(p\right)\right) \\ &= p \end{aligned} \tag{9}$$

for all $p \in [0, 1]$.

According to Equations (7), (8) and (9), we obtain for the distribution of P in case of a two tailed test: $F_P(p|H_0) = 2 \cdot \frac{p}{2} = p$. Note that we divide the probability p to equal parts between both tails. Therefore, the random variable P is uniformly distributed on the interval $[0, 1]$ when H_0 is true. □

Figures 1 and 2 show the histograms for simulated p-values under the null hypothesis and the alternative hypothesis respectively. The p-values are generated by the Kolmogorov-Smirnov test which is carried out over and over again for the problem of testing test whether or not data are coming from a standard normal distribution.

In the case when alternative hypothesis is true, the data are generated by a normal distribution with expected value 0.05 and standard deviation 1. Altogether, 100 different runs are made for data samples of length 1000. Figure 1 confirms that the p-values follow a uniform distribution on the unit interval $[0, 1]$ when the null hypothesis is true.

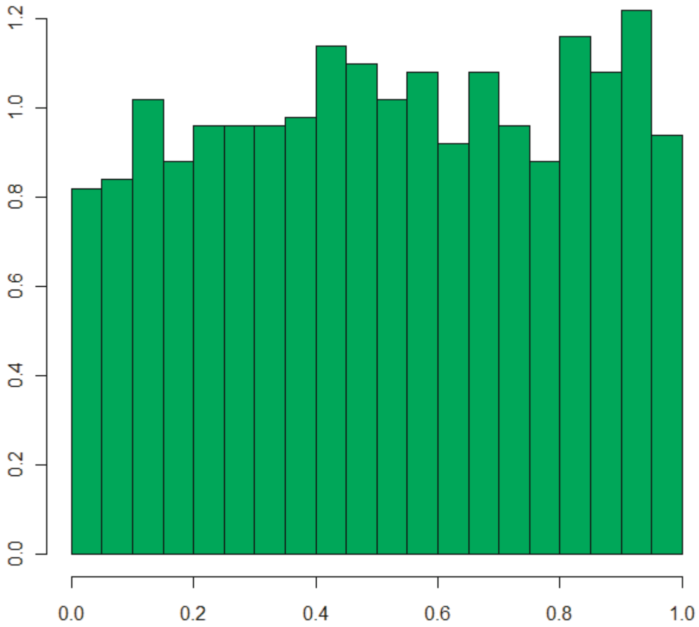

Fig. 1 Histogram for p-values under the null hypothesis.

The histogram in Figure 2 for the alternative hypothesis shows a different situation. The sampling distribution here is clearly not uniform anymore, the majority of values is close to zero and the amount of values decreases towards the p-value one.

Thus, we are interested in the question: how are p-values distributed when the alternative hypothesis holds? The distribution is given by Equation (10) (see [4]).

$$\begin{aligned} F_P(p|H_1) &= \Pr\left(P \leq p|H_1\right) \\ &= \Pr\left(1 - F_T\left(t\right) \leq p|H_1\right) \\ &= \Pr\left(F_T\left(t\right) \geq \left(1-p\right)|H_1\right) \\ &= 1 - \Pr\left(F_T\left(t\right) \leq \left(1-p\right)|H_1\right) \\ &= 1 - G_T\left(F_T^{-1}\left(1-p\right)\right) \end{aligned} \tag{10}$$

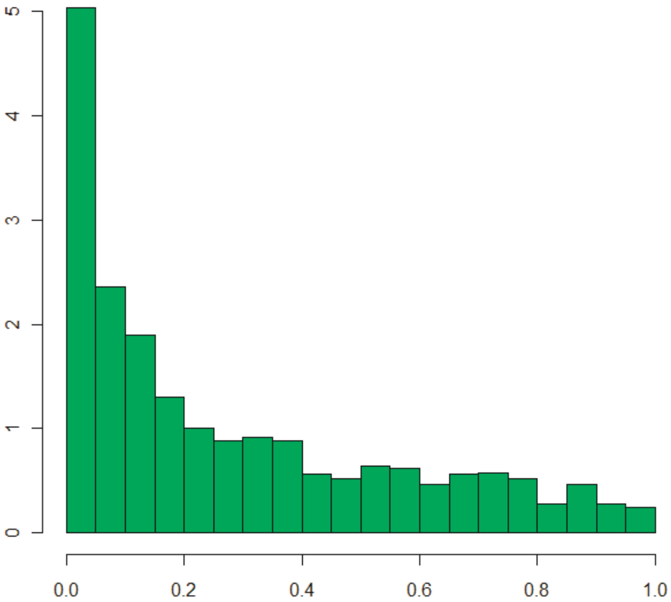

Fig. 2 Histogram for p-values under the alternative hypothesis.

where G_T is the distribution of the test statistic T under the alternative hypothesis. Here we only consider upper-tailed one-sided tests. As Equation (10) shows, the distribution of the p-values in this case depends on the test statistic distribution under H_0 as well as under H_1 hypothesis.

Hence, knowing the distribution of p-values under both hypotheses, a meta analysis can be performed. Since for each alternative hypothesis – in most cases the alternative is a composite hypothesis representing not a single but a set of distributions – and therefore for each G_T the distribution of p-values under H_1 is different (see Equation (10)) we restrict further considerations to the uniformity of p-values under H_0.

The most obvious way to carry out a meta analysis is to perform a goodness of fit test on the obtained p-values during multiple testing. For instance, the Kolmogorov-Smirnov test (an implementation is available in the R statistics library [3]) can be used for that purpose. However, the following problem should be taken into account: in order to carry out a meta analysis of p-values, neither a sliding window nor an incremental computation can be used for change detection. Indeed, the general assumption for hypothesis tests that the considered random variables are independent and identically distributed (i.i.d.) does not hold for overlapping sliding windows. By the application of sliding windows or incremental computation the next p-value is highly dependent on the previous ones. The reason for this problem is that almost the same values are used by the hypothesis test, correspondingly the computed neighbouring p-values would be approximately equal. Therefore, for this approach only non-overlapping windows should be used during change detection. As a consequence, we can not use the comparison between data from an earlier window with data from newer one when an abrupt change occurs, since in such a case H_0 would be false only once and therefore only one p-value would not come from a uniform distribution. Nevertheless, this approach shows good results when a test is used in order to proof whether the data follow a known or assumed distribution or to detect drift in the data generating process.

4 Experimental Results

Our approach has been implemented in Java using R-libraries and has been tested with artificial data. For the data generation process the following model was used: first n_1 time points data are generated from a standard normal distribution, i.e. $X_i \sim N(0, 1)$ for $i \in \{1, \ldots, n_1\}$. At time point $n_1 + 1$ a change occurs and the data are normally distributed with the following settings: $\mu = 0.1$ and $\sigma = 1$, i.e. $N(0.1, 1)$.

Our meta analysis of p-values has been applied to this data set. The Kolmogorov-Smirnov test for standard normality of the data was carried out for non-overlapping time windows. The size of the window for the change detection was chosen to be 500. Afterwards, the sliding window of size 100 was used for the meta analysis of the obtained p-values. In order to test the distribution of the p-values a Kolmogorov-Smirnov test for uniformity is used. A meta p-value is consequently the result of this test. Figure 3 illustrates described technique.

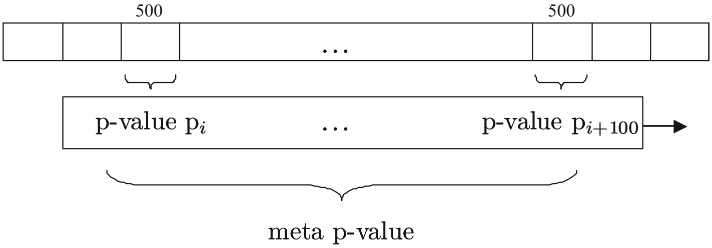

Fig. 3 Two windows for change detection.

The change occurred at the time point 59489. The computed p-values for this part of the data are as follows:

After the change occurs, the null hypothesis can be rejected (depending on the chosen α). However, from time to time H_0 cannot be rejected. Furthermore, for some parts without change H_0 is erroneously rejected, even

Table 1 p-values obtained during change detection.

time window	p-value
[57000; 57499]	0.07486618716777954
[57500; 57999]	0.1401818038913014
[58000; 58499]	0.9941898244705528
[58500; 58999]	0.9291249862216258
[59000; 59499]	0.5020298421810007
[59500; 59999]	0.01168733233091191
[60000; 60499]	0.05625967117647695
[60500; 60999]	0.6789664978854166
[61000; 61499]	0.394486208210243
[61500; 61999]	0.05360718854238174
[62000; 62499]	0.7747463214977733

though the underlying distribution did not change at that time. For instance for the interval [45500; 45999] the p-value is 0.018673 and consequently H_0 can be rejected for all $\alpha \geq 0.020$. Whereas as Table 2 shows, all meta p-values are smaller than 0.05 starting from the window [41000; 65999] and all meta p-values before are larger than 0.05.

Table 2 Meta p-values obtained during change detection.

time window	meta p-value
[40000; 64999]	0.1599219
[40500; 65499]	0.0809654
[41000; 65999]	0.0377086
[41500; 66499]	0.0161466
[42000; 66999]	0.0063506
[42500; 67499]	0.0022917

For the next example the data were generated as follows:

$$Y_t = \sum_{i=1}^{t} |X_i| \,. \tag{11}$$

We assume the random variables X_i to be normally distributed with expected value $\mu = 0$ and variance σ_1^2, i.e. $X_i \sim N\left(0,\ \sigma_1^2\right)$. To make the situation more realistic, we consider the following model:

$$Z_t \sim N\left(y_t, \sigma_2^2\right). \tag{12}$$

The process (12) can be understood as a constant model with drift and noise. The noise follows a normal distribution whose expected value equals the actual value of the random walk and whose variance is σ_2^2. The data were generated with the following parameters: $\sigma_1 = 0.00000008, \sigma_2 = 0.002$. Figure 4 shows the generated data.

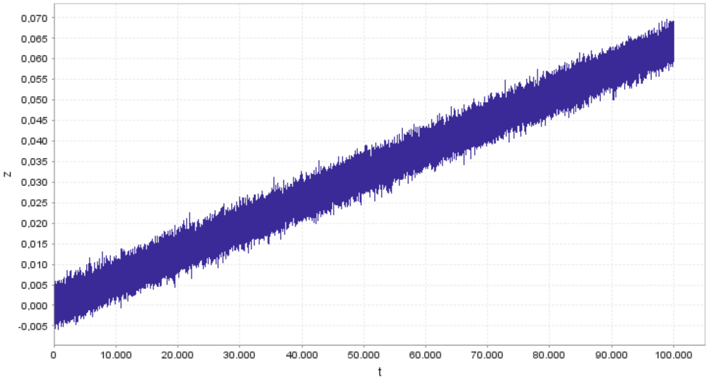

Fig. 4 Two windows for change detection.

In order to detect changes, the two sample t-test was applied to this data set. In such a way we can test whether the data from the old and new windows have the same mean. Two non-overlapping windows of size 500 are used. For the meta analysis, similar as before, the Kolmogorov-Smirnov test for uniformity is applied to a sliding window of size 50. Since the mean changes very slightly, sometimes H_0 can not be rejected (depending on the chosen α), as can be seen from Table 3, whereas all meta p-values provide the strong evidence that the data is non-stationary (all obtained meta p-values are smaller than 10^{-9}).

As Tables 1, 2 and 3 show, the meta p-values are more reliable than p-values. However, it should be taken into account that more time is needed until a change can be detected. Therefore, this approach is not suitable when very fast reaction to the occurred change is required. Whereas when more attention is paid to the accuracy of change detection, meta p-values provide a good solution to the problem of multiple testing for non-stationarity of the data. For instance such kind of change detection can be used for changes caused by slow wear and abrasion of materials, here the fast reaction is not required but the information about the speed of wear.

Table 3 p-values obtained during change detection.

time window t	time window $t+1$	p-value
[0; 499]	[500; 999]	0.04019542802527917
[500; 999]	[1000; 1499]	0.01835982391226245
[1000; 1499]	[1500; 1999]	0.02694198995888841
[1500; 1999]	[2000; 2499]	0.00590771301502357
[2000; 2499]	[2500; 2999]	0.00000051742252253
[2500; 2999]	[3000; 3499]	0.21019670543166669
[3000; 3499]	[3500; 3999]	0.02610004716763162
[3500; 3999]	[4000; 4499]	0.01388767893804595
[4000; 4499]	[4500; 4999]	0.02986063639554551
[4500; 4999]	[5000; 5499]	0.00174724618341983
[5000; 5499]	[5500; 5999]	0.21651140022620408
[5500; 5999]	[6000; 6499]	0.00180512145155431

5 Conclusion

Change detection is a crucial aspect for non-stationary data streams or "evolving systems". It has been demonstrated in [11] that naïve adaption without taking any effort to distinguish between noise and true changes of the underlying sample distribution can lead to very undesired results. Statistical measures and tests can help to discover true changes in the distribution and to distinguish them from random noise. However, the following problem arises: when a test is carried out over and over again, the probability to erroneously rejecting the null hypothesis increases with the amount of applied tests. In this work, we have discussed the problem of multiple testing during change detection and proposed classical methods as well as a new approach to cope with the multiple testing issue.

Bonferroni correction and the Bonferroni-Holm method adjust the significance level α in order to correct the occurrence of incorrect rejections of H_0 leading to a very conservative approach that will seldom indicate a change in the data stream. Our proposed approach is based on the uniformity of the p-values under the null hypothesis. In such a way, not only the p-values but also the meta p-values are taken into account by the change detection. This approach shows good results even in cases where Bonferroni correction and the Bonferroni-Holm method could not achieve any improvement. Although we have only considered the distribution of the p-values under the null hypothesis, it could be useful to study the distribution of p-values under the alternative hypothesis, too.

References

1. Basseville M, Nikiforov I (1993) *Detection of Abrupt Changes: Theory and Application.* Prentice Hall Prentice Hall, Upper Saddle River, New Jersey (1993)
2. Bhattacharya B, Habtzghi D (2002) Median of the p value under the alternative hypothesis. *The American Statistician* 56:202–206
3. Crawley M (2005) *Statistics: An Introduction using R.* J. Wiley & Sons, New York
4. Donahue RMJ (1999) A note on information seldom reported via the P value. *The American Statistician* 53(4):303–306
5. Gibbons J, Pratt J (1975) p-values: Interpretation and methodology. *The American Statistician* 29:20–25
6. Gustafsson F (2000) Adaptive Filtering and Change Detection. J. Wiley & SOns, New York
7. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70
8. Sackrowitz H, Samuel-Cahn E (1999) p-values as random variables—expected p-values. *The American Statistician* 53(4):326–331
9. Shaffer JP (1995) Multiple hypothesis testing. *Ann. Rev. Psych* 46:561–584
10. Sheskin D (1997) *Handbook of Parametric and Nonparametric Statistical Procedures.* CRC-Press, Boca Raton
11. Tschumitschew K, Klawonn F (2010) The need for benchmarks with data from stochastic processes and meta-models in evolving systems. *Int. Symp. Evolving Intelligent Systems*, 30–33. SSAISB, Leicester
12. Tschumitschew K, Klawonn F (2012) Incremental statistical measures. In: Sayed-Mouchaweh M, Lughofer E (eds.) *Learning in non-stationary environments: Methods and Applications*, Chap. 2. Springer, New York