Identifying features with concept drift in multidimensional data using statistical tests

Piotr Sobolewski and Michał Woźniak

Wrocław University of Technology
ul. Wybrzeże Wyspiańskiego 27
50-370 Wrocław
{piotr.sobolewski,michal.wozniak}@pwr.edu.pl

Abstract. Concept drift is a common problem in the data streams, which makes the classifiers no longer valid. In the multidimensional data, this problem becomes difficult to tackle. This paper examines the possibilities of identifying the specific features, in which concept drift occurs. This allows to limit the scope of the necessary update in the classification system. As a tool, we select a popular Kolmogorov-Smirnov test statistic.

Keywords: Concept drift, detection, statistical test

1 Introduction

Due to the evolution of the internet and the expansion of the decision making technology, the systems designed for classifying the data streams [2] recently became a popular area of research. In the field of machine learning, data streams are defined as sources of continuous data generation, examples of which can be found in real life e.g., shopping trends, stock market, weather control, surveillance systems or health care. Classification task in these areas is often hindered by various factors which cause undesirable changes in the data classification rules. Such phenomenon is called concept drift [6] and it is a major problem in the classification systems.

There are various methods described in the machine learning literature for defending against concept drift, mostly deploying one of the two popular strategies [4]:

- Adapting a learner at the regular intervals without considering whether the changes have really occurred or not,
- First detecting the concept changes and then adapting a learner to them.

The idea presented in this article has a potential of improving the classifier adaptation process as well as enhancing the efficiency of the concept drift detection algorithms.

2 Problem description

We assume that in the multidimensional data, concept drift may influence only some specific features, leaving all other features in the same conceptual distribution model. Identifying these features may improve the adaptation of the classification systems, as well as provide useful information for the sophisticated concept drift detection algorithms, such as LDCnet [9].

Our previous experiments [8][10] have shown, that the popular test statistics, such as the Kolmogorov-Smirnov test [7], Wilcoxon rank sum [11] or Wald-Wolfowitz test [12] are capable of detecting concept drift with a similar efficiency as advanced methods, designed specifically for this purpose, such as the CNF test [3]. The most efficient in our experiments was the Kolomogorov-Smirnov test statistic, therefore we have selected it for further analysis in this article.

Kolomogorov-Smirnov test is a non-parametric statistic, as it makes no assumption about the distribution of data and therefore can be deployed on any data.

For the two-sample test, a Kolmogorov-Smirnov statistic is computed as

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{1}$$

where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of samples computed as:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1\{x_i \leq t\}, \tag{2}$$

where $(x_1, ..., x_n)$ are independent and identically distributed (i.i.d.) random variables laying in the real numbers domain with a common cumulative distribution function. The statistic is used to perform a KS-test to reject the null hypothesis at level α by computing:

$$\sqrt{\frac{nm}{n_m}} D_{n,m} > K_\alpha, \tag{3}$$

where K_α calculated from:

$$Pr(K \leq K_\alpha) = 1 - \alpha, \tag{4}$$

and K is a Kolmogorov distribution computed as:

$$K = \sup_{t \in [0,1]} |B(t)|, \tag{5}$$

$B(t)$ being the Brownian bridge [5].

In short, the Kolmogorov-Smirnov test compares the distributions of two samples by measuring a distance between the empirical distribution functions, taking into account both their location and shape.

In this paper we evaluate the possibilities of applying the Kolmogorov-Smirnov test statistic as a tool for identifying the fueatures, which are influenced by concept drift. For this purpose, the tool needs to accurately classify the true positives (sensitivity) as well as the true negatives (specificity).

3 Data

Due to limited availability of the real data with concept drift, the data used in experiments is taken from the UCI Repository of datasets [1] and concept drift is simulated by swapping the features with each other.

In mathematical notation, if a reference dataset DS is characterized by n features f then the concept drift is applied by swapping any two features i and j with each other. Data with swapped features forms a new dataset $DS_{i,j}$ and the role of the algorithm is to identify which features are influenced by concept drift (i.e. find the i and j).

Example swap of features 1 and 2:

$$\begin{array}{c} DS = [f_1, f_2, ..., f_n] \\ DS_{1,2} = [f_2, f_1, ..., f_n] \end{array}, \tag{6}$$

The swaps are made for each possible pair of features, resulting in $\binom{n}{2}$ combinations, where n is the dimensionality of the dataset. Each of the dataset is described in general by the number of samples and number of features in Tab. 1.

Table 1. Datasets

Dataset	# of features	# of samples
breast	9	683
credit-australian	14	690
haberman	3	306
heart-c	13	297
heart-statlog	13	270
ionosphere	34	351
kr-vs-kp	36	3196
letter-recognition	16	20000
mfeat-mor	6	2000
nursery	8	12960
optdigits	64	3823
page-blocks	10	5473
pendigits	16	7494
pima-indians-diabetes	8	768
segmentation	19	210
tic-tac-toe	9	958
vehicle	18	846
vote	16	232
waveform	21	5000
yeast	8	1484

This method of simulating concept drift is relatively common in the machine learning literature [13].

4 Experiments

In the experiments, we use the original dataset D as the reference data and the drifted datasets $D_{i,j}$ (i and j indicate the features which are swapped), with the samples randomly drawn from the datasets $D_{i,j}$ and groupped into data windows DW of various sizes s.

The Kolomogorov-Smirnov statistic is evaluated on every feature f in the data window to reject the null hypothesis that the values arise from the same population as the values of features in the reference dataset D with confidence level of 5%. It means, that if the test returns the p-value lower than 0.05, then the analyzed feature is considered to be influenced by concept drift and the detection signal is noted and added to the scores.

In order to evaluate the specificity, i.e. the ability to identify the true negatives, the data windows which do not include any feature swap are evaluated and if in this test the statistic returns the p-value lower than 0.05, then the algorithm makes a mistake, as it results in a false positive concept drift detection.

A short description of the experimental process is described in the pseudocode in Fig. 1.

Algorithm 1 Pseudo-code of a single loop in experimental series

```
Notations:
  DS - original dataset with n features,
  DS_{i,j} - dataset with swapped features i and j,
  DW^s_{i,j} - data window of size s with features i and j swapped,
Single loop of experiment seriers:
  DW^s = draw s random data samples from DS
  For i = 1 to f
    For j = 1 to f
      DW^s_{i,j} = swap features i and j in DW^s
      For k = 1 to f
        Evaluate KS statistic on feature k of DW^s_{i,j} and feature k of D
        IF p-value < 0.05
          Note concept drift for feature k
        END IF
      END FOR
    END FOR
  END FOR
```

In the presented way, the sensitivity and specificity of the Kolmogorov-Smirnov test statistic are evaluated for every possible feature swap and for various sizes of the data windows.

5 Results

All presented values are averaged from the series of 1000 trials.

Tables 2 and 3 show the percentage of correctly detected concept drifts in certain features for the *breast* dataset (size of data window 20 and 50, respectively), where columns are the base features and rows are the features which swap them. The diagonals are the percentage of detected false positives, the lower the value the higher the specificity of the algorithm.

Tables 4 and 5 show how the window size influences the performance of algorithm for the *breast* and $credit - australian$ datasets. The tables store the results obtained by swapping the first feature with other features. The results obtained for the *breast* dataset are also presented on the Diagram 1 for a more clear view of the efficiency trend in the domain of window size.

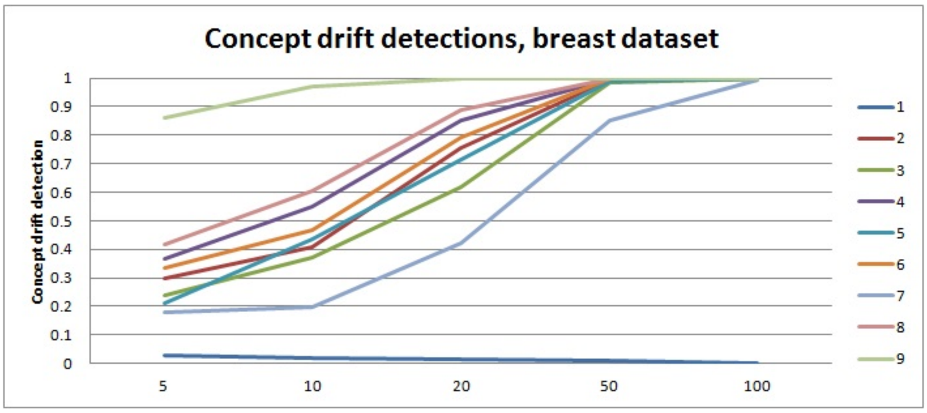

Fig. 1. Concept drift detection ratio for various window sizes.

Finally, Tab. 6 shows the overall performance of the Kolmogorov-Smirnov statistic in identifying the features affected by concept drift, divided into the sensitivity and specificity scores for each of the datasets and for various window sizes. Specificity is presented only for window size 20, as the results were not significantly different for other window sizes.

Table 2. Concept drift detection ratio in breast dataset, window size = 20

win size 20	Base feature								
Swap feature	1	2	3	4	5	6	7	8	9
1	0.02	0.7	0.67	0.86	0.82	0.89	0.27	0.91	1
2	0.65	0.02	0	0.02	0.93	0.01	0.5	0	0.47
3	0.59	0.01	0.01	0.03	0.93	0.03	0.43	0.07	0.63
4	0.84	0.02	0.02	0	0.95	0.01	0.71	0.03	0.34
5	0.68	1	0.99	0.99	0	1	0.03	1	1
6	0.73	0.02	0.08	0.02	1	0	0.8	0.04	0.33
7	0.39	0.56	0.62	0.63	0.17	0.78	0.03	0.85	1
8	0.89	0.03	0.03	0.01	1	0.02	0.83	0	0.16
9	1	0.6	0.58	0.27	1	0.42	0.99	0.13	0

Table 3. Concept drift detection ratio in breast dataset, window size = 20

win size 20	Base feature								
Swap feature	1	2	3	4	5	6	7	8	9
1	0.02	1	0.98	1	1	1	0.79	1	1
2	1	0	0	0.02	1	0.07	1	0.03	0.92
3	0.99	0.01	0	0.1	1	0.08	0.95	0.12	0.99
4	1	0.03	0.05	0.01	1	0.05	0.99	0.03	0.83
5	1	1	1	1	0	1	0.48	1	1
6	1	0.1	0.11	0.14	1	0	1	0.11	0.89
7	0.88	0.98	0.96	1	0.57	1	0	1	1
8	1	0.07	0.09	0.01	1	0.03	1	0.01	0.59
9	1	0.99	1	0.89	1	0.97	1	0.58	0

Table 4. Window size influence, *breast* dataset.

breast	Feature 1 swapped with feature..								
win size	1	2	3	4	5	6	7	8	9
5	0.03	0.30	0.24	0.37	0.21	0.34	0.18	0.42	0.86
10	0.02	0.41	0.37	0.55	0.43	0.47	0.20	0.60	0.97
20	0.01	0.75	0.62	0.85	0.72	0.79	0.42	0.89	1
50	0.01	0.99	0.99	1.00	0.99	1.00	0.85	1.00	1
100	0.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1

Table 5. Window size influence, $credit - australian$ dataset.

credit-aus	Feature 1 swapped with feature..													
win size	1	2	3	4	5	6	7	8	9	10	11	12	13	14
5	0.00	1.00	0.89	0.91	1.00	0.99	0.46	0.02	0.06	0.26	0.06	1.00	0.96	0.63
10	0.00	1.00	0.98	0.98	1.00	1.00	0.62	0.04	0.13	0.32	0.09	1.00	1.00	0.78
20	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.09	0.31	0.90	0.24	1.00	1.00	1.00
50	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.83	1.00	0.70	1.00	1.00	1.00
100	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.63	0.99	1.00	0.97	1.00	1.00	1.00

Table 6. Overall sensitivity and specificity scores.

	Sensitivity								Specificity	
win. size	20		50		100		200		20	
Dataset	avg	var	avg	var	avg	var	avg	var	avg	var
breast	0.53	0.15	0.70	0.17	0.77	0.14	0.82	0.11	0.99	0.00
credit-australian	0.92	0.06	0.95	0.04	0.96	0.03	0.97	0.02	0.99	0.00
haberman	0.92	0.06	0.95	0.04	0.96	0.03	0.97	0.02	0.99	0.00
heart-c	0.94	0.04	0.98	0.01	1.00	0.00	1.00	0.00	0.99	0.00
heart-statlog	0.95	0.03	0.98	0.01	1.00	0.00	1.00	0.00	0.99	0.00
ionosphere	0.60	0.16	0.70	0.15	0.76	0.14	0.81	0.13	0.97	0.00
kr-vs-kp	0.41	0.18	0.56	0.20	0.65	0.19	0.72	0.18	1.00	0.00
letter-recognition	0.48	0.18	0.62	0.19	0.72	0.17	0.78	0.15	0.99	0.00
mfeat-mor	0.48	0.19	0.62	0.19	0.72	0.17	0.78	0.15	0.99	0.00
nursery	0.47	0.18	0.62	0.19	0.71	0.18	0.77	0.16	0.99	0.00
optdigits	0.69	0.16	0.78	0.13	0.84	0.11	0.89	0.08	0.98	0.00
page-blocks	0.69	0.16	0.79	0.13	0.84	0.10	0.89	0.08	0.98	0.00
pendigits	0.68	0.16	0.79	0.13	0.84	0.10	0.89	0.08	0.97	0.00
pima-indians-diabetes	0.69	0.16	0.79	0.13	0.84	0.10	0.89	0.08	0.97	0.00
segmentation	0.70	0.16	0.79	0.13	0.84	0.10	0.89	0.08	0.98	0.00
tic-tac-toe	0.68	0.17	0.78	0.14	0.83	0.11	0.87	0.09	0.98	0.00
vehicle	0.71	0.16	0.80	0.13	0.85	0.10	0.89	0.08	0.98	0.00
vote	0.66	0.17	0.75	0.15	0.81	0.12	0.86	0.10	0.98	0.00
waveform	0.68	0.16	0.78	0.13	0.83	0.11	0.88	0.08	0.97	0.00
yeast	0.68	0.16	0.78	0.13	0.83	0.11	0.88	0.08	0.97	0.00

6 Discussion

In this paper we have proposed an unsupervised tool for enhancing the methods coping with concept drift. We have evaluated the efficiency of the Kolmogorov-Smirnov test statistic in detecting the features affected by concept drift in the multidimensional data.

The most apparent conclusion is that the performance of algorithm depends on the data window size. Fig. 1 clearly shows this relation.

Regardless of the window size, algorithm achieves a very high specificity score, proving that the tool performs very well with true negatives, i.e. when there is no drift. It means, that the tool can be used for detecting features with concept drift without the need to worry about the false positive detections.

On the other hand, sensitivity i.e. the true positive detection rate, leaves a field for improvement. With increasing window size, sensitivity of the tool also increases, what suggests that the tool is more feasible for problems, which do not require a very limited window size.

Overall, the performance of the proposed tool is on a decent level, as e.g. in the optdigits dataset scenario, which has 2^{64} possible feature swap combinations, algorithm correctly identifies on average 88% of them with only 8% variance. Pairing it with the fact that the method does not require any supervision, the Kolmogorov-Smirnov test statistic can be considered an efficient tool for detecting the features with concept drift in multidimensional data. This functionality may be used for supporting the adaptation of classifiers as well as improving algorithms designed for detecting concept drift, such as LDCnet [9].

Further research aims on expanding the functionality of the mentioned LDCnet algorithm using the presented technique to battle concept drift in the multidimensional data.

References

1. D.J. Newman A. Asuncion. UCI machine learning repository, 2007.
2. Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '02, pages 1–16, New York, NY, USA, 2002. ACM.
3. Anton Dries and Ulrich Rückert. Adaptive concept drift detection. *Stat. Anal. Data Min.*, 2(56):311–327, December 2009.
4. Russell Greiner, Adam J. Grove, and Dan Roth. Learning cost-sensitive active classifiers. *Artif. Intell.*, 139(2):137–174, August 2002.
5. Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion (Grundlehren der mathematischen Wissenschaften)*. Springer, 3rd edition, December 2004.
6. Jeffrey C. Schlimmer and Richard H. Granger, Jr. Incremental learning from noisy data. *Mach. Learn.*, 1(3):317–354, March 1986.
7. N. V. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.*, 19:279–281, 1948.
8. P. Sobolewski and M. Wozniak. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.
9. P. Sobolewski and M. Wozniak. Ldcnet: minimizing the cost of supervision for various types of concept drift. In *Proceedings of the CIDUE 2013 - IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments*, CIDUE '13, pages 68–75, 2013.
10. Piotr Sobolewski and Michal Wozniak. Comparable study of statistical tests for virtual concept drift detection. In Robert Burduk, Konrad Jackowski, Marek Kurzynski, Michal Wozniak, and Andrzej Zolnierek, editors, *CORES*, volume 226 of *Advances in Intelligent Systems and Computing*, pages 329–337. Springer, 2013.
11. Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

12. J. Wolfowitz. On Wald's Proof of the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20:601–602, 1949.
13. Indre Zliobaite and Ludmila I. Kuncheva. Determining the training window for small sample size classification with concept drift. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 447–452, Washington, DC, USA, 2009. IEEE Computer Society.