Concept Drift Detection in Prediction Classifiers for Determining Gender in Metabolomics Analysis

A. Kostoval[1], J. Schwarzerova[1,2]

[1]Department of Biomedical Engineering Faculty of Electrical Engineering and Communication, Brno University of Technology Brno, Czech Republic

[2]Molecular Systems Biology (MOSYS), University of Vienna, Vienna, Austria

E-mail: 221515@vut.cz, Jana.Schwarzerova@vut.cz

Abstract—Currently, one of the most challenges in data analysis is connected to prediction modeling including dynamic information. Metabolomics analysis focuses on data presented dynamic information in real-time such as time-series data. Unfortunately, prediction models based on time series data are often affected by a phenomenon called concept drift. This phenomenon can reduce the accuracy of prediction models which is an unwanted effect. On the other hand, concept drift analysis can be useful in finding confounding factors. This study is divided into two parts. The first part presents the modeling of prediction classifiers based on metabolite data. The second part of this study brings concept drift detection in the created classified models. This study presented approaches to identify one of the confounding factors in human biology.

Keywords—Concept drift, Concept drift detection, Metabolomics, Machine learning, Prediction modeling

1. INTRODUCTION

The concept drift is defined as unexpected changes between input and output data. Sometimes concept drift occurs when test sets are changing unpredictably. Subsequently, the model is unable to respond correctly because training sets were different. It follows, that concept drift negatively affects prediction models which were being trained on dynamically changing data.

Concept drift analysis studies the detection and correction of an unwanted phenomenon that is caused by obvious changes in real-time [1]. These changes are characterized in data distribution and can be detected in a prediction model [2]. Thus, the main goal of concept drift detection focuses to bring algorithms that can detect changes in data distribution. As a result of concept drift detection, we would ensure the long-term accuracy, reliability, and stability of prediction models.

Metabolomics is one of omics science focused on detection, quantification, and analysis of metabolites in an organism [3]. Metabolites are small molecules, smaller than 1500 *Da*. Metabolomics and the study of prediction models based on metabolites data leads to earlier detection of human disease, which is reflected in metabolomes such as diabetes mellitus, cancer, etc. [3]. Nevertheless, prediction models based on metabolite data are affected by the concept drift phenomenon [4]. Therefore, this concept drift phenomenon is necessary to detect and reveal confounding factors that are hidden.

2. MATERIALS AND DATA

In this study, data were taken from the cohort study by Chu et al. [5]. Metabolomics datasets include 534 healthy subjects (237 males and 296 females) in ages 18 – 75. The study brought two different datasets which were measured using different techniques. The first dataset represented platform Brainshake Metabolomics (BM) which was measured by principle nuclear magnetic resonance and included 231 features with 200 absolute concentrations. The second dataset represented platform General Metabolomics (GM) based on flow injection TOF-M. GM included 1586 features with 257 absolute concentrations. These data are freely available at: https://hfgp.bbmri.nl/.

3. METHODS

The whole methodology of this study is divided into three parts. The first step includes pre-processing phasis focused on data preparation. The second part is focused on model prediction classifiers using different approaches. The last part is connected to concept drift analysis. In the last part, the concept drift detection was performed, and the confounding factor was identified.

Firstly, datasets were divided into inputs and targets data. Inputs data represent concentrations of metabolites. Targets data classify phenotype of gender into two binary classes (male – 0, female – 1). Furthermore, both datasets were split by 10-fold cross-validation to estimate classify skill of prediction models. Training and testing datasets were divided into a ratio of 9:1 and cross-validation ran 10 times.

Secondly, the prediction classifiers were modeled. The classifiers were trained to determine the gender of a patient based on measured metabolite concentration. Prediction models were implemented in Python using the library Scikit-learn [6]. Namely, we used: Logistic Regression (LR), Gradient Boosting (GB), Random Forest (RF), and Naïve Bayes (NB). All these methods were applied to model classifiers.

The last step of our methodology included concept drift detection. The concept drift detectors were used on metabolomics models. The concept drift principle is shown in Figure 1. Firstly, input data are predicted by a decision-making process. In the next step, the concept drift detector detects deflection in data distribution, detector warns possibility of concept drift presence. Nevertheless, concept drift detectors have two thresholds. The first limit is warning but the second limit announces exact detection of change in data distribution. Finally, these detected changes are appropriately revised and again predicted by a decision-making process.

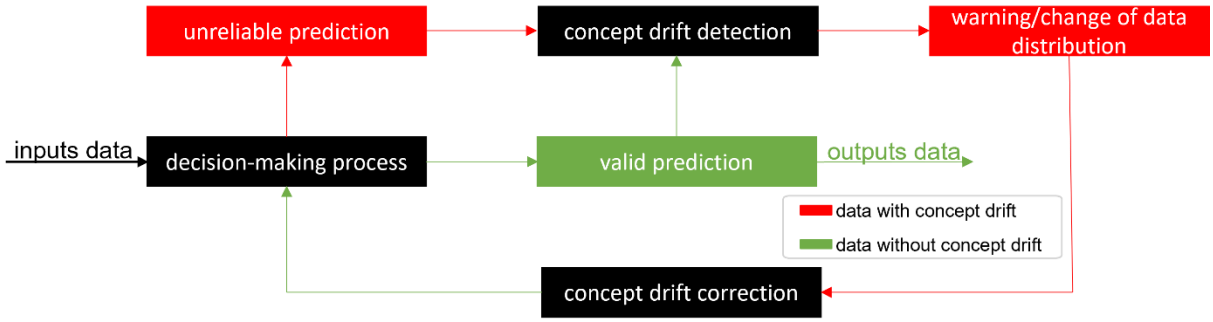

Figure 1: Pipeline of methodology for concept drift detection and correction

The first method which was used in our study is Drift Detection Method (DDM) [3]. DDM is based on the monitor error rate of the classification, which is understood as the probability of incorrect prediction. The second method is the Early Drift Detection Method (EDDM) [3]. EDDM is very similar to DDM, but EDDM analyzes also changes in distance between two consecutive misclassified objects. EDDM is useful to detect gradual changes in data distribution. On the other hand, DDM is better for detecting sudden concept drift. DDM and EDDM were implemented from Scikit-multiflow [7] for each created classifier model.

4. RESULTS AND DISCUSSION

The accuracy of our prediction models was estimated using 10-fold cross-validation. Table I shows evaluation parameters presented final accuracy for each method. The highest value of accuracy connects to the GB method, see Table I. Table II presents another metric for evaluating the accuracy of the classification was the F1-score. This metric compares the prediction of classification models with the target. The best accuracy and F1-score achieved the LR model in the case of the BM dataset and the GB model in the case of the GM dataset. On the other hand, the lowest value of accuracy is identified in prediction models based on the NB approach and F1-score values confirm this statement.

Table I: The accuracy (dimensionless quantity) values of classification models for datasets BM/GM

	LR	**GB**	**RF**	**NB**
BM	0,85	0,83	0,79	0,69
GM	0,85	0,88	0,84	0,78

Table II: The F1-score (dimensionless quantity) values of classification models for datasets BM/GM

	LR	GB	RF	NB
BM	0,81	0,81	0,79	0,73
GM	0,83	0,89	0,87	0,76

The concept drift detectors bring revealing of concept drift in our created models. Table III shows the number of concept drift detection for models created by BM data. Table IV includes the number of concept drift detection in models based on GM data.

Table III: Number of warnings (W) and changes (CH) detection in data distribution in BM dataset

	LR	GB	RF	NB
DDM	35 W / 0 CH	22 W / 0 CH	39 W / 0 CH	17 W / 0 CH
EDDM	0 W / 1 CH	0 W / 1 CH	0 W / 1 CH	0 W / 1 CH

Table IV: Number of warnings (W) and changes (CH) detection in data distribution in GM dataset

	LR	GB	RF	NB
DDM	0 W / 0 CH	0 W / 0 CH	0 W / 0 CH	0 W / 0 CH
EDDM	5 W / 0 CH	0 W / 1 CH	18 W / 0 CH	5 W / 0 CH

Precisely, Figure 2 shows the plot for comparison number of concept drift detection using methods DDM and EDDM.

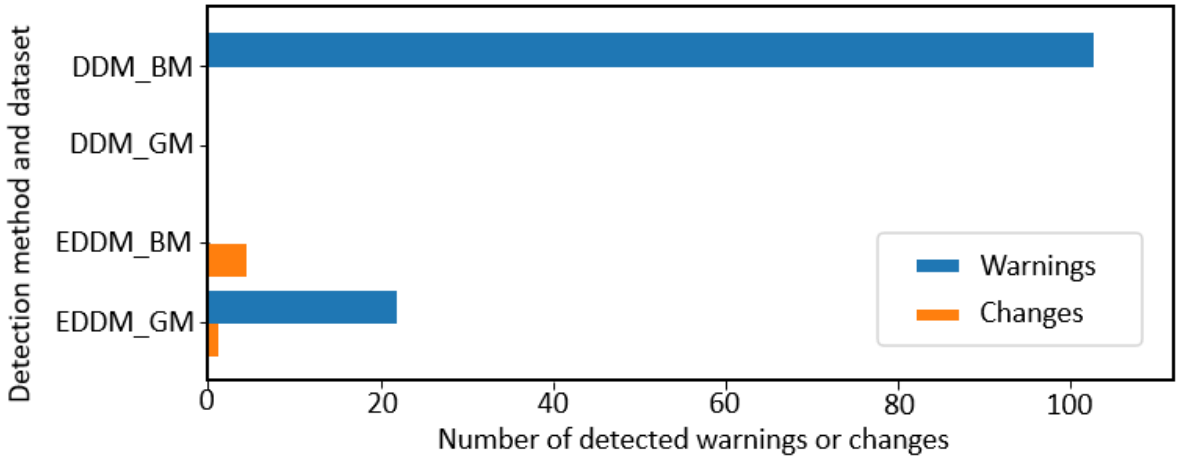

Figure 2: Summary concept drifts by detection methods DDM and EDDM. DDM_BM (EDDM_BM) is the sum of concept drifts detection by DDM (EDDM) for each classifier model based on BM data. Similarly, DDM_GM (EDDM_GM) is the sum of concept drift detection for classifier models based on GM data.

The highest number of warning levels was detected by DDM in RF models. In each of the RF models, data distribution includes significant changes detected as concept drift phenomenon. EDDM detector revealed more concept drift detection than DDM. Therefore, EDDM is more suitable for detecting the concept drift in metabolomic prediction, which corresponds with the findings in the study [4]. In the end, we identified concept drift according to patient age which is a promising factor as a confounding factor in the human metabolomics analysis. Figure 3 shows the detection of concept drift occurring in the adolescent period. Regarding it, the confounding factor presented the age of a patient is identified. Thus, our study brings confirmation revealed confounding factor from the study [4].

Figure 3: Visualization of concept drifts detection according to the age of patients

5. CONCLUSION

Metabolomics brings a new era allowing to deal with prediction phenotype based on metabolites concentrations. However, metabolomics focuses on data presented dynamic information in real-time such as time-series data. Thus, prediction models based on metabolite data can be included to concept drift which reduces prediction accuracy. This study brings 8 prediction classifiers that predict gender based on metabolites concentrations. The main message of the study is to reveal an innovative view on metabolomics analysis with detecting confounding factors like is concept drift.

In created classification models, the concept drift was detected using DDM and EDDM. The EDDM is more appropriate for concept drift detection in metabolomic prediction than DDM. Thanks to concept drift detection, confounding factor related metabolomics analysis was identified as the age of a patient. This finding will help to create more accurate models for the early diagnosis, which is essential to a full recovery, or economically less demanding treatment.

ACKNOWLEDGMENT

This work has been supported by grant FEKT-K-21-6878 realised within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69 /0.0/0.0/19_073/ 0016948, which is financed from the OP RDE.

REFERENCES

[1] WEBB, Geoffrey I., et al. Characterizing concept drift. Data Mining and Knowledge Discovery, 2016, 30.4: 964-994.

[2] YU, Shujian, et al. Concept drift detection and adaptation with hierarchical hypothesis testing. Journal of the Franklin Institute, 2019, 356.5: 3187-3215.

[3] IDLE, Jeffrey R.; GONZALEZ, Frank J. Metabolomics. Cell metabolism, 2007, 6.5: 348-351.

[4] SCHWARZEROVA, Jana, et al. An Innovative Perspective on Metabolomics Data Analysis in Biomedical Research Using Concept Drift Detection. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021. p. 3075-3082.

[5] CHU, Xiaojing, et al. Integration of metabolomics, genomics, and immune phenotypes reveals the causal roles of metabolites in disease. Genome biology, 2021, 22.1: 1-22.

[6] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[7] MONTIEL Jacob et al., "Scikit-multiflow: A multi-output streaming framework", The Journal of Machine Learning Research, vol. 19, no. 1, pp. 2915-2914, 2018.