Diversity in Ensemble Model for Classification of Data Streams with Concept Drift

Michal Kolárik, Martin Sarnovský, Ján Paralič
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics,
Technical University Košice, Letna 9
04001 Kosice, Slovakia
michal.kolarik@tuke.sk, martin.sarnovsky@tuke.sk, jan.paralic@tuke.sk,

2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI) | 978-1-7281-8053-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/SAMI50585.2021.9378625

***Abstract*—Data streams can be defined as the continuous stream of data in many forms coming from different sources. Data streams are usually non-stationary with continually changing their underlying structure. Solving of predictive or classification tasks on such data must consider this aspect. Traditional machine learning models applied on the drifting data may become invalid in the case when a concept change appears. To tackle this problem, we must utilize special adaptive learning models, which utilize various tools able to reflect the drifting data. One of the most popular groups of such methods are adaptive ensembles. This paper describes the work focused on the design and implementation of a novel adaptive ensemble learning model, which is based on the construction of a robust ensemble consisting of a heterogeneous set of its members. We used k-NN, Naive Bayes and Hoeffding trees as base learners and implemented an update mechanism, which considers dynamic class-weighting and Q statistics diversity calculation to ensure the diversity of the ensemble. The model was experimentally evaluated on the streaming datasets, and the effects of the diversity calculation were analyzed.**

***Index Terms*—data streams, concept drift, ensemble learning, ensemble diversity**

I. Introduction

In recent years, demand for the processing of the data streams increased. Data are often collected from many devices and services in a continuous fashion, often in high velocity and therefore require a different set of technologies and methods than the static data. Streams are often dynamic; the data in the stream is evolving, which means that the concepts described by such a data may change in the process. Such changes in the non-stationary streams are called concept drift. When solving predictive or classification tasks on the streaming data where the data generation process is not stationary, we must use the methods capable of handling such dynamic data. One of the most important features of such models is the ability to adapt to the new incoming data and their continuous incorporation in the model. Multiple advanced machine learning algorithms have been developed to meet those requirements, including active learning methods, adaptive models with concept drift detectors or ensemble methods. Ensemble methods currently present a popular group of methods for classification on the drifting data. Ensembles proved to be very useful and often superior to other methods on the static data and the fact that the model consists of multiple members, makes it a good candidate to handle the changing concepts. Adaptive ensembles are usually based on a continuous re-training and updating of the ensemble members using the new data, with various mechanisms of the class-weighting and the members voting implemented.

This paper describes a new heterogeneous ensemble method, which combines different base learner models within the ensemble and during the update of the ensemble members also utilizes diversity metrics of the ensemble members to ensure the robustness of the entire ensemble. The paper is organized as follows: Section II presents the background and fundamental concepts in the field of the data streams processing and the drifting data. Following section Section III briefly summarizes the measurements of diversity used in ensemble models. Section IV presents the designed ensemble method. The following Section V outlines the datasets used in the experiments and their results. Finally, Section VI summarizes the main contributions of the proposed method and sketches some future research possibilities.

II. Background and related work

There are several definitions of data streams in the literature. In general, a data stream is an unbound, ordered sequence of data elements [1]. The stream elements appear from its source continuously, over time. Data streams may differ, in the format of the data elements, in the time interval between the elements of the stream, or the size of particular stream items. Most of the streams are generated at high speeds, with the stream elements generating rapidly, while its size is rather small. There are two kinds of streams; stationary and non-stationary. Stationary data streams are composed of the elements, which data distribution is rather constant, while non-stationary data streams are evolving. There are also several restrictions and special requirement for data stream processing [2]: accuracy should not differ significantly from batch methods, timeliness should correspond to the interval of data arrival and adaptivity should be a key part in data stream processing because of its dynamic characteristic. In data streams, the data distribution changes during the time, as stream flows, which usually affects the target concepts (classes in classification tasks).

Concept drift, in general, represents the change of the underlying data distribution, which results in changes of the target concepts [3]. There are two types of concept drift:

- Real drift – a change in the data stream, which results in the changes of the target concept;
- Virtual drift – a change in the data stream, which necessarily does not affect the target class boundaries.

Drifts may appear in the stream with various velocities and frequencies. According to when and how the drifts appear in the stream [4], we distinguish between incremental, gradual, sudden (or abrupt) and reoccurring concept drifts.

When performing classification on the drifting streams, we must use the predictive models able to reflect the changes in the data and to adapt to a changed concept. Some machine learning models are naturally incremental (online), such as Naive Bayes. Others such as decision trees require dynamic structure changes for performing adaptive processing. Ensemble models proved suitable for a wide range of classification tasks on the static data. Ensembles work on the assumption that a collection of weak independent classifiers performs better when they are working together than performance of random model. Such composed model consists of a collection of its members (also called base learners or experts), and for classification of the unknown examples, it combines their individual decisions into the final one. There are several types of popular ensemble models frequently used in different predictive tasks, e.g., Bagging [5], Boosting [6] or Random Forests [7]. Ensembles are also very frequently used in classification on drifting data. There are multiple streaming modifications of bagging and boosting methods available, successfully applied to the drifting data. Very popular are Online Boosting [8] and Online Bagging [9]. Multiple variations of such methods were also described, combining them with the drift detectors [10], with added modifications to improve the performance on imbalanced data [11] or applied on large-scale data streams [12]. In general, most of these methods rely on the homogeneous ensemble, consisting of members of the same type (same base learner algorithm). The difference is mostly in the voting mechanism or different class-weighting. There are also some models with heterogenous approaches [13] [14].Presented work focuses on the creation of a heterogeneous ensemble, consisting of various models trained using different base learners. On top of that, we include also the specific metrics to measure the diversity of particular members in the ensemble. Such a metric is used in the phase of the model adaptation to the new data when the worse performing ensemble members are updated by the new members, which are diverse from the existing ones. The following section briefly describes the diversity measures used in the ensemble models and design of the ensemble model with the diversity measurement we implemented in the update phase.

III. Diversity in ensemble models

Diversity in ensemble models is a critical requirement when combining the ensemble experts. Diversity can be achieved using three main strategies [15]: block-based data, weighting-data, or filtering-data. In spite of that, in our work we use a different approach by combining heterogeneous experts based on diversity measurements of particular models. However, measuring diversity is ambiguous, because there is no generally accepted definition of the diversity of the ensemble. Kunecheva and Whitaker [16] describe the ten best-known methods for calculating diversity of experts in the ensemble model. There are two basic ways of measuring the diversity:

- Pairwise diversity measures – Q statistics, The correlation coefficient ρ, The disagreement measure, The double-fault measure;
- Non-pairwise diversity measures – The entropy measure E, Kohavi-Wolpert variance, Measurement of interrater agreement κ, The measure of "difficulty" θ, Generalized diversity, Coincident failure diversity.

In this research, Q statistic pairwise diversity was used because of its implementation, calculation and interpretation of independence. Values of Q statistic varies between -1 and 1 where positive values belong to classifiers that tend to classify the same samples correctly, and negative values belong to classifiers with error on different instances. Higher diversity is achieved at lower Q statistic value.

IV. DDCW ensemble method

Following the discussion of diversity in ensemble models in Section III, we propose a new ensemble method called Diversified Dynamic Class Weighted (DDCW). The proposed algorithm is inspired by Dynamic Weighted Majority (DWM) algorithm [17] but improves its majority voting mechanism (scoring vector) by extending it to set of vectors with weights not only for individual experts but also for the particular target classes. This creates a matrix of weights for experts to every target class. The size of this matrix is dynamically changing based on the number of experts and the number of actual target classes. The individual weights are updated dynamically in multiple ways at every period of processing the data. The model decisions are based on the highest score, which means the sum of weights of all experts to the current target class. This allows the model to better adapt to unbalanced multi-class data. In other words, each of the experts can focus on a different target class. Besides that, the experts in the ensemble model are weighted not only by correct predictions but also based on measuring Q statistic diversity in the model. The more diverse models are preferred. To achieve higher diversity between experts, we decided to use a heterogeneous ensemble model. Experts are dynamically selected from the set of online experts consisting of Hoeffding Tree, k-NN and Naive Bayes. The size of the model is dynamically changing based on the current model performance. The DDCW model can be categorized as a passive online model because it does not contain any drift detection method. Despite that model adapts to a concept drift by adding of the new experts to the ensemble and replacing the weak ones.

As depicted in Figure 1, the model is composed of m ensemble experts, randomly selected from a set of experts including Hoeffding Tree, Naive Bayes and k-NN classifiers. Then, for each expert E_i and for each known target class c the weights are updated so that each target class has a sum of weights equals 1. In the beginning, the weights are set to

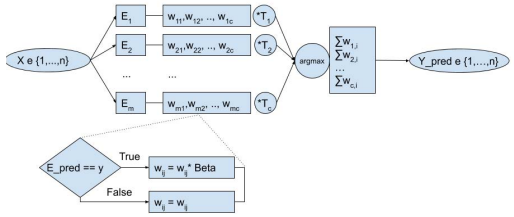

Fig. 1. Overall scheme of the proposed ensemble model.

equal values. After each period model updates experts' weights w_{ic}. The updates of weights are based on experts accuracy in each class (weight of correct expert in the correct target class is multiplied by Beta-parameter). Then, the weights of each expert are multiplied by the life-time coefficient T_i (to prefer newer experts against older ones) and updated by its pair diversity value. According to Section III lower Q statistic value means higher diversity of the model. Thus, model updates the weights in an inverse manner $w_{ij} + (1 - diversity\_value)$. Q statistic diversity is calculated between each pair of experts based on their predictions in the actual period. After updates, weights are normalized so that the sum of weight for each class equals 1. For each sample, the contribution from each of the experts is calculated per each target class, and the final decision is made on the basis of the highest score.

V. Experiments

In our experiments we focused on observing the impact of diversity on the model performance. Therefore, several experiments were performed on datasets Electricity, KDD99, LED and Stagger using DDCW models with different period parameters. In addition to accuracy and F1, the monitored metrics were also the state of particular experts and the value of their diversity Q statistics.

A. Datasets Description

For the following experiments, it was necessary to select suitable input data that would meet the required characteristics. Selected datasets are from real-world collections as well as synthetically generated data. Datasets are diverse, used in different types of classification task (multi-class classification, binary classification), or with a different drift type present. In the case of real datasets, the distribution of the target class is much more unbalanced. Selected datasets are summarized in the Table I.

TABLE I
Datasets used in the experiments.

Dataset	Dataset type	Drift type	Samples	Features	Classes
Eletricity	real	?	45 312	8	2
KDD99_10%	real	?	489 000	41	23
Stagger	synthetic	abrupt	100 000	3	2
LED	synthetic	gradual	100 000	24	10

- Electricity [18] - contains 45,312 samples described using 8 attributes. The dataset contains data collected every 30 minutes during two years and it was provided by the Australian Society of New South Wales Electricity. The classification task represents an increase (up) or decrease (down) of electricity prices.
- KDD99 [19] - well-known dataset in the processing of data streams. It is suitable for testing the precision of models because it contains data samples (normal behaviour or different tzpes of attacks) that record behavior of the observed system for only a short time of the recorded events. The dataset represents the detection of cyber-attacks on a server where different types of attacks occur. For these experiments, the selected dataset sample (10%) was used. The dataset contains 489,000 samples, 41 attributes and 23 target classes.
- Stagger [20] - a synthetic dataset with an abrupt type of concept drift. It contains 3 nominal attributes: size, colour and shape. For the experiment, the dataset contains 100,000 samples where drift appears after every 33,333 samples with a width of 50 samples to simulate the drift. 10% of samples may represent the noise.
- LED [21] - is a synthetic dataset with gradual concept drift. The goal is to predict a digital number on the 7-segment display. Each number is displayed with a probability of 10%. The dataset contains 7 relevant attributes and 17 irrelevant ones. Drift appears every 25,000 samples with a width of 500 samples. There is also 10% the noise data.

In the first series of experiments, observations were made on the Electricity dataset. In this set of experiments, we wanted to compare the impact of diversity in the DDCW model. DDCW model with enabled diversity to update weights (DDCW-1E) with DDCW model where the diversity of ensembles had a minimum impact on updating weights (DDCW-1D). The progress of the performance on Electricity dataset is shown in Figure 2.

The results proved, that the DDCW model with diversity mechanism enabled achieving approximately 1% better accuracy and F1 measure than DDCW model without taken into account the diversity. However, the reasons for the influence of model with diversity over the model without diversity on accuracy or F1 are not directly observable from the visualization of the model progress. There are some points in visualization when the performance of the model with diversity (DDCW-1E) did not drop down as in model without diversity (DDCW-1D).

We also tried to identify internal models' behaviour. Thus, we looked at the progress of diversity in both models DDCW-1E and DDCW-1D when processing the Electricity dataset and the result is also depicted in Figure 3. We can see the lower values of diversity in the DDCW-1E model, but as we mentioned in Section III., the lower values of our diversity measure (Q statistic) means higher model diversity and experts within this model are a little bit more diverse than the experts in the DDCW-1D model. The more diverse models should have better performance.

Figure 4 depicts the number of experts in the ensemble. DDCW model is dynamic, which means that the number

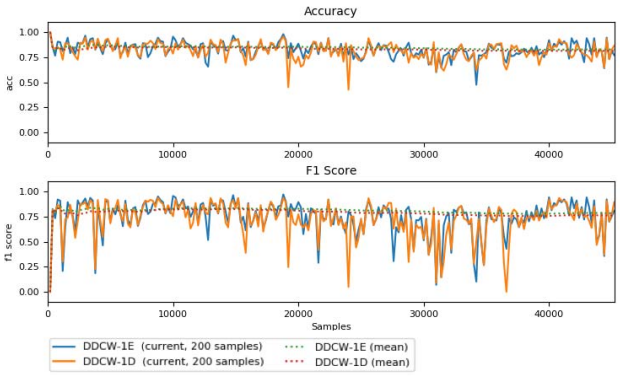

Fig. 2. Progress of accuracy and F1 in DDCW model with and without diversity on dataset Electricity.

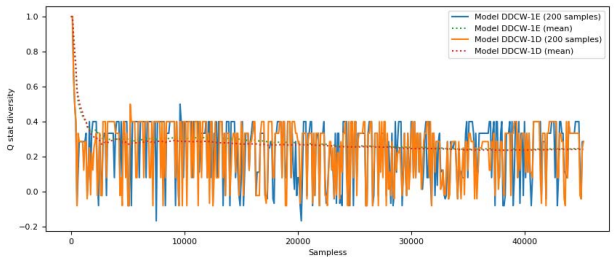

Fig. 3. Progress of diversity measure Q in DDCW model with and without diversity on dataset Electricity.

of ensemble members varies over the time. In both models, we can see that the DDCW model relies on a relatively smaller ensembles, where its members are being updated more frequently. But, the average number of experts in models is lower for the DDCW-1E model type. This means that the model with diversity requires a lower number of experts whereas the performance of this model is higher.

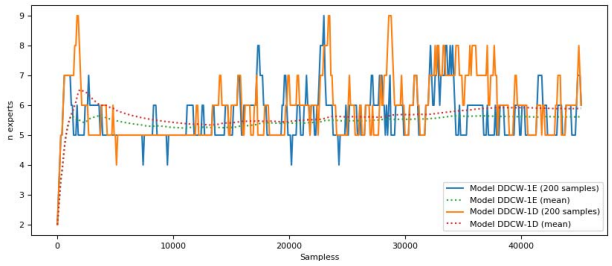

Fig. 4. Number of experts in DDCW model with and without diversity on dataset Electricity.

Complete results of our experiments are listed in Table II. The table summarizes the performance of both types of models for all 4 datasets. Models with enabled diversity, i.e. DDCW-1E models (in the table with Div flag) and traditional ensembles, i.e. DDCW-1D in different settings for period values (parameter p). The period p defines, how many samples of the stream fit into one chunk and are processed at once. Parameter p also influences the nominal size of the values for calculating the model diversity and then updating the experts weights.

Experiments proved that smaller values of the chunk size lead to better performance and in general, we can observe, that the model with diversity measurement (Div flag) brings slight improvement of the models in both, accuracy and F1 metrics. In some experiments, for example, with the LED dataset, the model with diversity initially achieved worse results, but with a period of 1000, it exceeds the model without enabled diversity. The higher period enables to calculate diversity from a bigger set of historical values and leads to model less prone to noise.

TABLE II
ACCURACY AND F1 OF DDCW MODEL WITH ENABLED AND DISABLED DIVERSITY USING DIFFERENT MODEL PERIODS.

	$p = 100$ Div	$p = 100$	$p = 500$ Div	$p = 500$	$p = 1000$ Div	$p = 1000$
dataset	acc/F1	acc/F1	acc/F1	acc/F1	acc/F1	acc/F1
Eletricity	**0.82/0.78**	0.81/0.77	**0.81/0.76**	0.79/0.73	**0.80/0.73**	0.79/0.72
KDD99_10%	0.99/0.64	0.99/0.64	0.99/0.63	0.99/0.63	0.99/0.63	0.99/0.63
Stagger	0.94/0.95	0.94/0.95	0.94/0.94	**0.95/0.95**	**0.94/0.95**	0.94/0.94
LED	0.85/0.85	**0.86/0.86**	0.87/0.87	0.87/0.87	**0.86/0.86**	0.84/0.84

In another series of experiments, the DDCW model was compared with other known ensemble models such as Accuracy Weighted Ensemble classifier (AWE) [22], DWM, OzaBagging (Oza) and Online Boosting classifier (OB). The results of this experiment are shown in Table III.

TABLE III
COMPARISON OF ACCURACY AND F1 METRICS OF EVALUATED ENSEMBLE MODELS.

	DDCW	DWM	AWE	OnlineBoosting	OzaBagging
dataset	acc/F1	acc/F1	acc/F1	acc/F1	acc/F1
Eletricity	**0.82/0.77**	0.81/0.76	0.77/0.70	0.79/0.75	0.78/0.74
KDD99_10%	**0.99/0.64**	0.98/0.51	0.43/0.05	0.99/0.64	0.99/0.63
Stagger	0.94/0.94	0.94/0.94	**0.95/0.95**	0.93/0.93	0.95/0.95
LED	0.84/0.84	0.83/0.34	**0.89/0.89**	0.85/0.85	0.84/0.84

VI. CONCLUSION

The work presented in this paper described the designed heterogeneous ensemble model DDCW with the implemented mechanism for ensuring a diversity of its members. The method utilizes various base learners, including Naive Bayes, k-NN and Hoeffding trees and employs dynamic class-weighting scheme and Q statistic to ensure a suitable level of diversity in the group of the ensemble members. The proposed model was experimentally evaluated on four datasets used in data streams classification tasks. We mostly focused on the implemented diversity measure and how it affects the model performance. On all of the studied datasets, our proposed DDCW model with the diversity mechanism enabled achieved slightly better results, both in terms of overall accuracy and F1 metric, leading to approximately 1% improvement in comparison with the traditional approach, without managing diversity of particular classifiers. When compared to some other popular ensemble methods, the DDCW model gained comparable results, being best in two out of the four datasets, proving that it is suitable for usage on the drifting data with

different types of concept drifts. In future work, we plan to explore the performance of the method in more in-depth experiments, evaluating it on a wider range of datasets and using a more fine-tuned hyperparameters.

ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-16-0213.

REFERENCES

[1] João Gama, Jesus Aguilar-Ruiz, and Ralf Klinkenberg. Knowledge discovery from data streams. *Intelligent Data Analysis*, 12(3):251–252, 2008.

[2] Viera Rozinajová, Anna Bou Ezzeddine, Gabriela Grmanová, Petra Vrablecová, and Miriama Pomffyová. Intelligent Analysis of Data Streams. *Towards Digital Intelligence Society: A Knowledge-based Approach*, pages 1–20, 2020.

[3] Indrė Žliobaitė, Mykola Pechenizkiy, and João Gama. An Overview of Concept Drift Applications. 2016.

[4] Indrė Žliobaitė. Learning under Concept Drift: an Overview. pages 1–36, 2010.

[5] Leo Breiman. Bagging predictors. *Machine Learning*, 1996.

[6] Yoav Freund and Robert E. Schapire. Experiments with a New Boosting Algorithm. *Proceedings of the 13th International Conference on Machine Learning*, 1996.

[7] Leo Breiman. Random forests. *Machine Learning*, 2001.

[8] Jan N. van Rijn, Geoffrey Holmes, Bernhard Pfahringer, and Joaquin Vanschoren. The online performance estimation framework: heterogeneous ensemble learning for data streams. *Machine Learning*, 2018.

[9] Nikunj C. Oza. Online bagging and boosting. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2005.

[10] Jesse Read, Albert Bifet, Geoff Holmes, and Bernhard Pfahringer. Scalable and efficient multi-label classification for evolving data streams. *Machine Learning*, 2012.

[11] Boyu Wang and Joelle Pineau. Online Bagging and Boosting for Imbalanced Data Streams. *IEEE Transactions on Knowledge and Data Engineering*, 2016.

[12] Y. Lv, S. Peng, Y. Yuan, C. Wang, P. Yin, J. Liu, and C. Wang. A classifier using online bagging ensemble method for big data stream learning. *Tsinghua Science and Technology*, 2019.

[13] Gabriela Grmanová, Peter Laurinec, Viera Rozinajová, Anna Bou Ezzeddine, Mária Lucká, Peter Lacko, Petra Vrablecová, and Pavol Návrat. Incremental ensemble learning for electricity load forecasting. *Acta Polytechnica Hungarica*, 13(2):97–117, 2016.

[14] Peng Zhang, Xingquan Zhu, Yong Shi, Li Guo, and Xindong Wu. Robust ensemble learning for mining noisy data streams. *Decision Support Systems*, 50(2):469 – 479, 2011.

[15] Imen Khamassi, Moamar Sayed-Mouchaweh, Moez Hammami, and Khaled Ghédira. A New Combination of Diversity Techniques in Ensemble Classifiers for Handling Complex Concept Drift. (August 2018):39–61, 2019.

[16] L.I. Kuncheva. Ten measures of diversity in classifier ensembles: limits for two classifiers. 2006.

[17] J. Zico Kolter and Marcus A. Maloof. Dynamic weighted majority: An ensemble method for drifting concepts. *Journal of Machine Learning Research*, 8:2755–2790, 2007.

[18] Michael Harries, U Nsw cse tr, and New South Wales. Splice-2 comparative evaluation: Electricity pricing. Technical report, 1999.

[19] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the KDD CUP 99 data set. In *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009*, 2009.

[20] Jeffrey C. Schlimmer and Richard H. Granger. Incremental Learning from Noisy Data. *Machine Learning*, 1986.

[21] A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. *Biometrics*, 1984.

[22] Dariusz Brzeziński and Jerzy Stefanowski. Accuracy updated ensemble for data streams with concept drift. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6679 LNAI(PART 2):155–163, 2011.