Faculty of Electrical Engineering and Informatics

Diversity in Ensemble Model for Classification of the 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 904001 Kosice, Slovakia

michal.kolarik@tuke.sk, martin.sarnovsky@tuke.sk, jan.paralic@tuke.sk

Contents

- Motivation
- Data stream definition
- Problems in processing data streams
- Concept drift types
- Models for data stream processing
- Proposed model
- Matrix of weights
- Comparing to other models
- Experiments
- Datasets
- Results
- Conclusion

Faculty of Electrical Engineering
and Informatics

Motivation of work

- Creation of new heterogeneous ensemble, consisting of various models trained using different base learners.
- We would like to include also the metrics to measure the diversity of the members in the ensemble.
- This metric could be used to select or improve members of model which are more diverse from other ones.

Faculty of Electrical Engineering and Informatics

Data stream definition

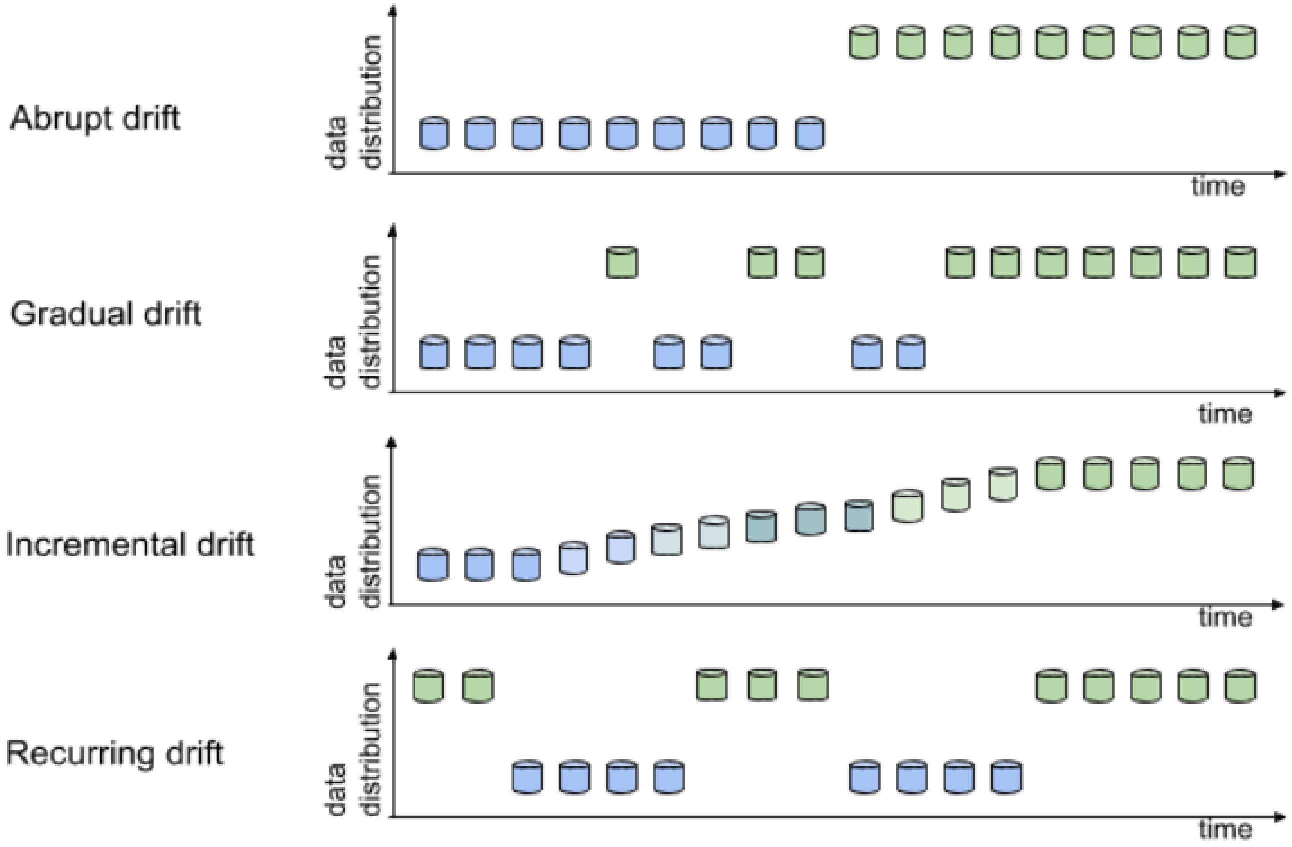
- Data stream is an unbound, ordered sequence of data elements.
- The stream elements appear from its source continuously, over time.
- Data streams may differ :
    - the format of the data elements
    - the time interval between the elements of the stream
    - the size of particular stream items
- Most of the streams are generated at high speeds.
- There are two kinds of streams: stationary and non-stationary.

Problems in processing data streams

- Limited resource for cpu, memory and time for fast predictions
    - Simpe, incremental or batch metod are used
- Phenomenon concept drift
    - Changes in data distribution in data streams
- Large volumes of data that are being continually produced at a high rate over time
- New data is constantly arriving even as the old data is being processed

Faculty of Electrical Engineering
and Informatics

Concept drift types



Types of concept drift

Models for data stream processing

Models with drift detection

- Main idea is to detect concept drift what triggers action in model structure:
  - Update of model classifiers
  - Actual classifiers set is replaced by new one
- Popular drift detectors:
  - DDM
  - EDDM
  - ADWIN
- Example models:
  - Diversity for Dealing with Drifts (DDD)
  - OnlineBoosting (Aboost)

Ensemble models

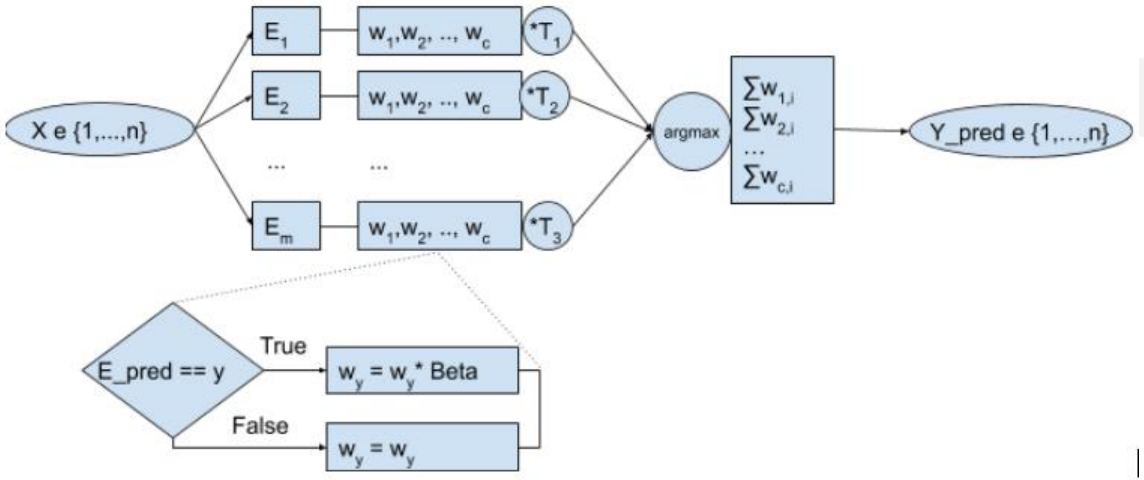
- Based on set of multiple simple classifiers
  - Ensemble create powerful model with diverse classifiers
  - This classifiers are updated or replaced dynamically based on model performance
  - Heterogeneous or homogeneous
- Final classification is based on voting rule
  - Majority voting, Weighted voting, stacking
- Popular models
  - Accuracy Weighted Ensemble (AWE)
  - Dynamic Weighted Majority (DWM)
  - Online Bagging (OzaBagging)

Faculty of Electrical Engineering and Informatics

Our proposed model

Diversified Dynamic Class Weighted (DDCW)

- Heterogenous ensemble model based on simple online classificators (experts)
    - Naive Bayes, KNN, Hoeffding Tree
- Experts score is decreased based on expert life in model – number of period
- After each period the experts weights are updated based on
    - expert accuracy
    - internal metric pair diversity Q statistic (between each pair of experts)
- The size of model is dynamic
    - experts are removed according its score
    - added when overall model performance is lower



Overall scheme of the proposed ensemble model.

Faculty of Electrical Engineering and Informatics

DDCW – matrix of weights

- Matrix of weights is used to store weight from each expert to each target class
  - Weights are dynamically updated based on expert performance
  - One expert can predict some classes very well but others poorly
  - This makes model suitable for unbalanced data streams
- Class score
  - Contribution of experts weight to target class
- Experts score
  - Sum of experts weights

	target class #1	target class #2	target class #3	experts score
expert #1	0,1	0,2	0,2	0,5
expert #2	0,4	0,6	0,1	1,1
expert #3	0,5	0,2	0,7	1,4
class score	0,5	0,0	0,7	

Table of weight matrix

Faculty of Electrical Engineering and Informatics

Comparting to other models

Proposed model DDCW

- Heterogeneous – based on different classifiers
- Incremental online model with small periods
- Updating weights after each period
- Matrix of weights
- Removing old and weak expert by new one
- Internal metrics for Q stat diversity

Other ensemble models

- Mostly homogeneous – one type of classifier
- Incremental or batch type with sliding window
- Voting based or weight based
- Different types of updating weights
- Dynamic or stable ensemble models

Experiments

- In first sets of experiment, we compare DDCW model with enabled and disabled model updates based on diversity
    - Goal was to determine impact of diversity to model behaviour and performance
- The other set of experiment was about comparsion DDCW model performance with other well know models
    - Similar dynamic models (AWE, DWM)
    - More robust models (OzaBagging, OnlineBoosting)

Datasets

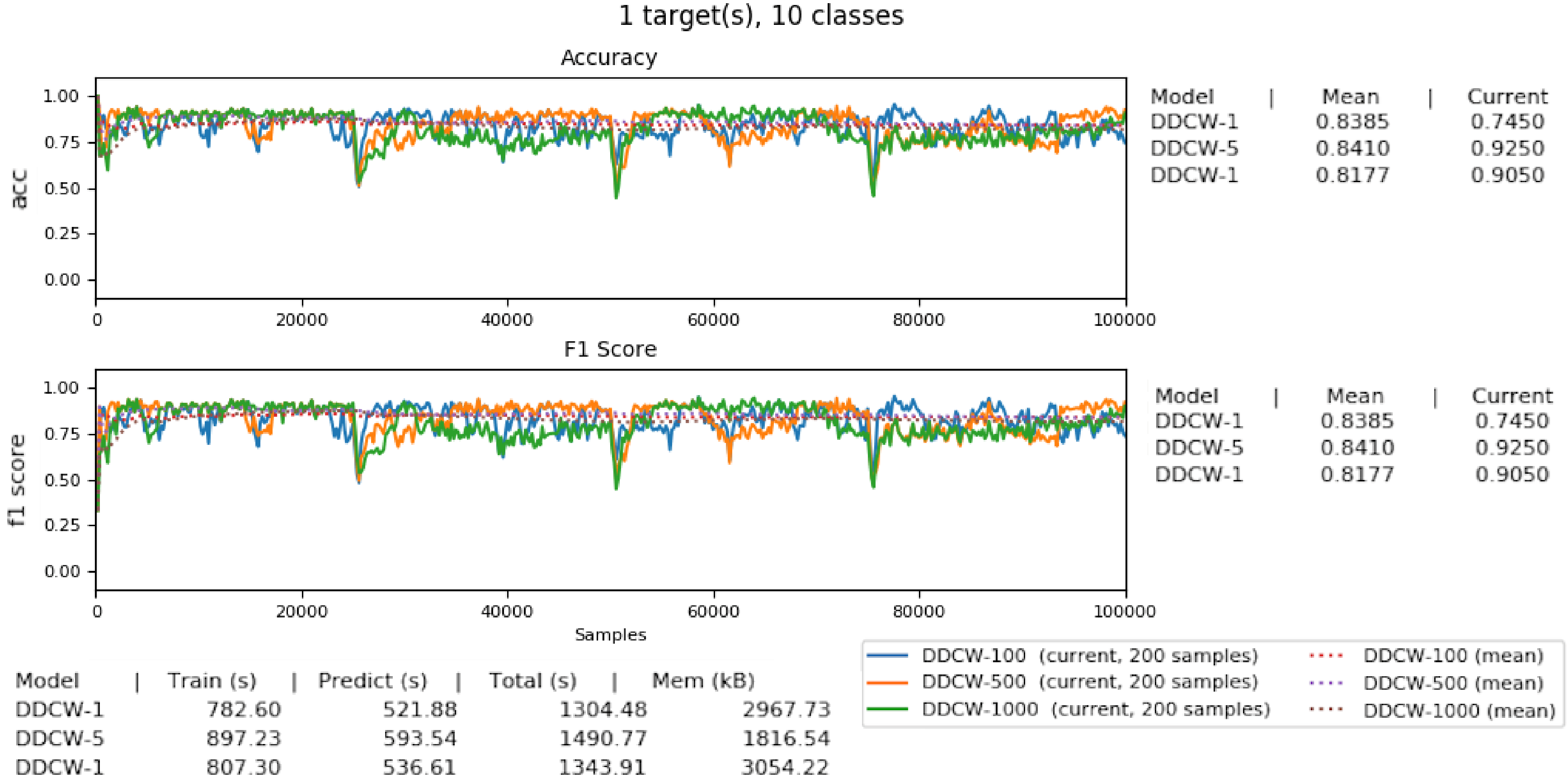
- Selected Datasets are from two categories:
  - Synthetic
  - Real
- Datasets covers tasks for binary and multiclass classification
- Datasets contains different types of concept drift
- We use **Gini Index** per windows of 1000 samples to visualize changes of concepts in datasets

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

Datasets used in the experiments.

Faculty of Electrical Engineering
and Informatics

Results of experiments with changing period size of model DDCW on dataset LED



Faculty of Electrical Engineering and Informatics

Progress of accuracy and f1 in DDCW model with and without diversity on dataset LED

Results of comparing models in table

dataset	$p = 100$ Div acc/F1	$p = 100$ acc/F1	$p = 500$ Div acc/F1	$p = 500$ acc/F1	$p = 1000$ Div acc/F1	$p = 1000$ 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

Accuracy and f1 of DDCW model with enabled and disabled diversity using different model periods.

dataset	DDCW acc/F1	DWM acc/F1	AWE acc/F1	OnlineBoosting acc/F1	OzaBagging 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

Comparsion of accuracy and f1 metrics of evaluated ensemble models.

Conclusion

- New heterogenous ensemble model DDCW with diversity mechanism
- Various base learners – Naive Bayes, k-NN and Hoeffding tree
- Class-weighted scheme and matrix of weights

- Various experiments on multiple datasets
- Model perform better compared to other tested models

- Future plans – experiments more in depth and wide range of datasets

Faculty of Electrical Engineering
and Informatics

Thank you