Leibniz Universität Hannover

A Decision-Based Dynamic Ensemble Selection Method for Concept Drift

Tyng-Jiun Kuo, Yaser Alagele

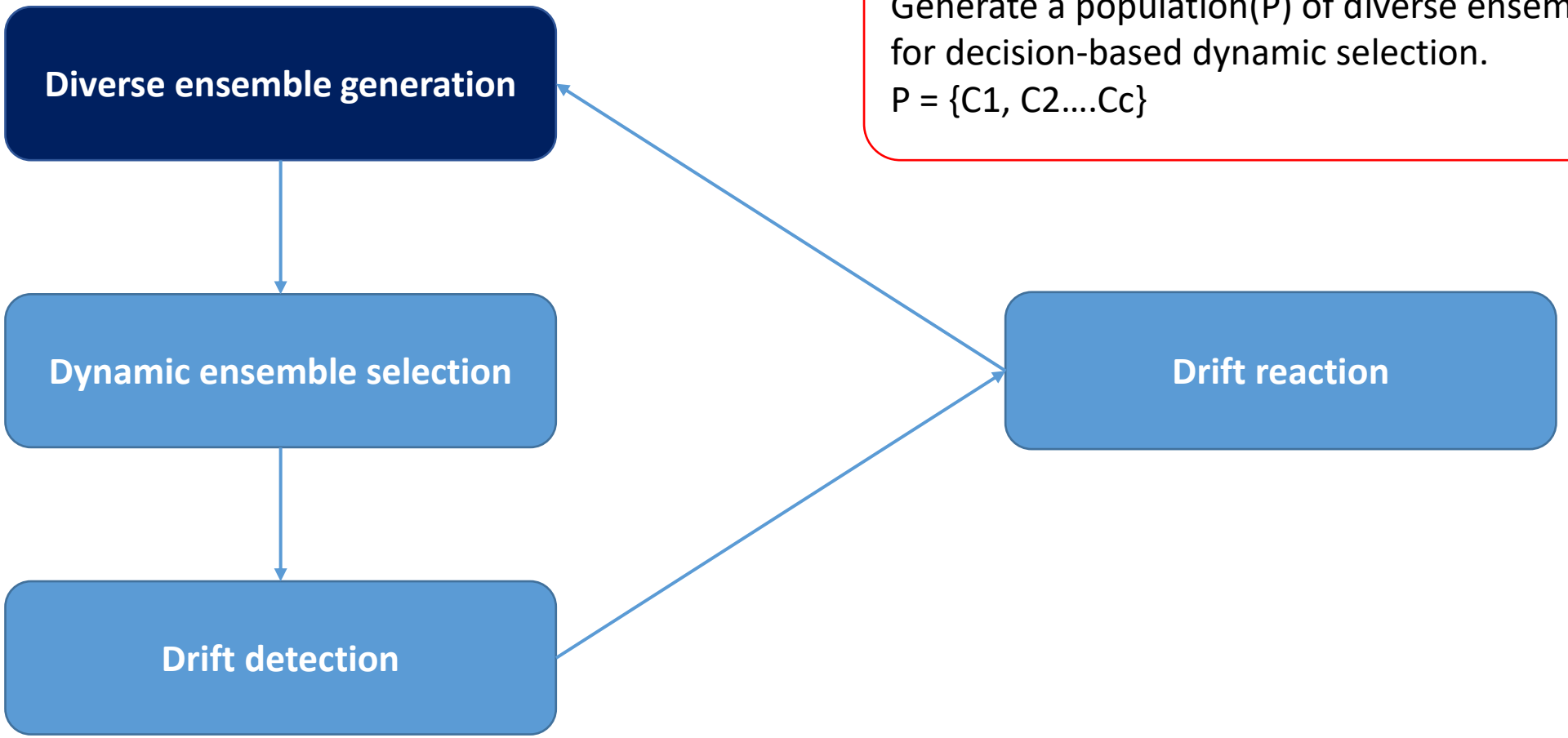
Introduction

- Problem: Concept drift
- when data are continuously generate in streams, data and target concepts may change over time.

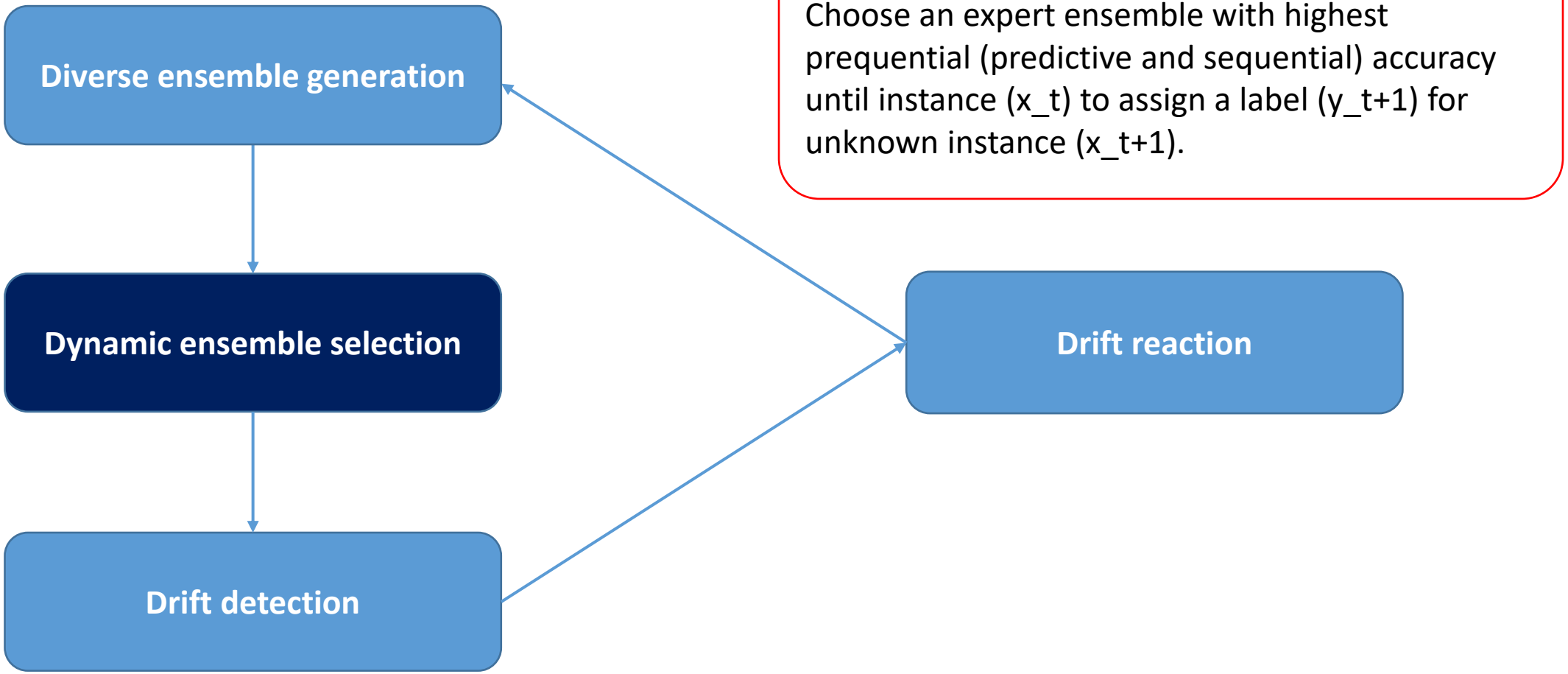
- Solution: Online method
- It focus on monitoring whether the class distribution is stable over time by observing the prediction of a classifier.

- The author proposed an decision-based online ensemble method Dynamic Ensemble Selection for Drift Detection (DESDD) for concept drift detection.

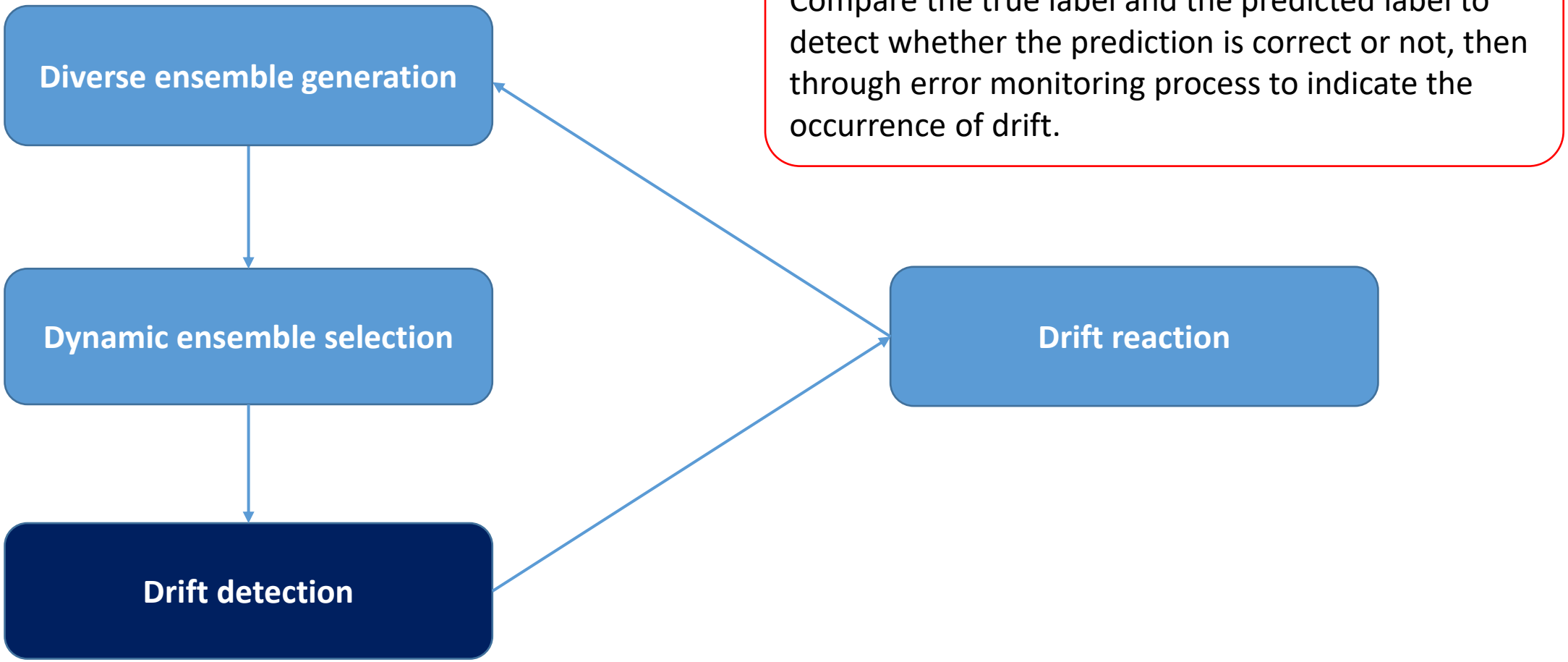
Method


Diverse ensemble generation
Dynamic ensemble selection
Drift detection
Drift reaction
Generate a population(P) of diverse ensemble(C) for decision-based dynamic selection.
P = {C1, C2….Cc}

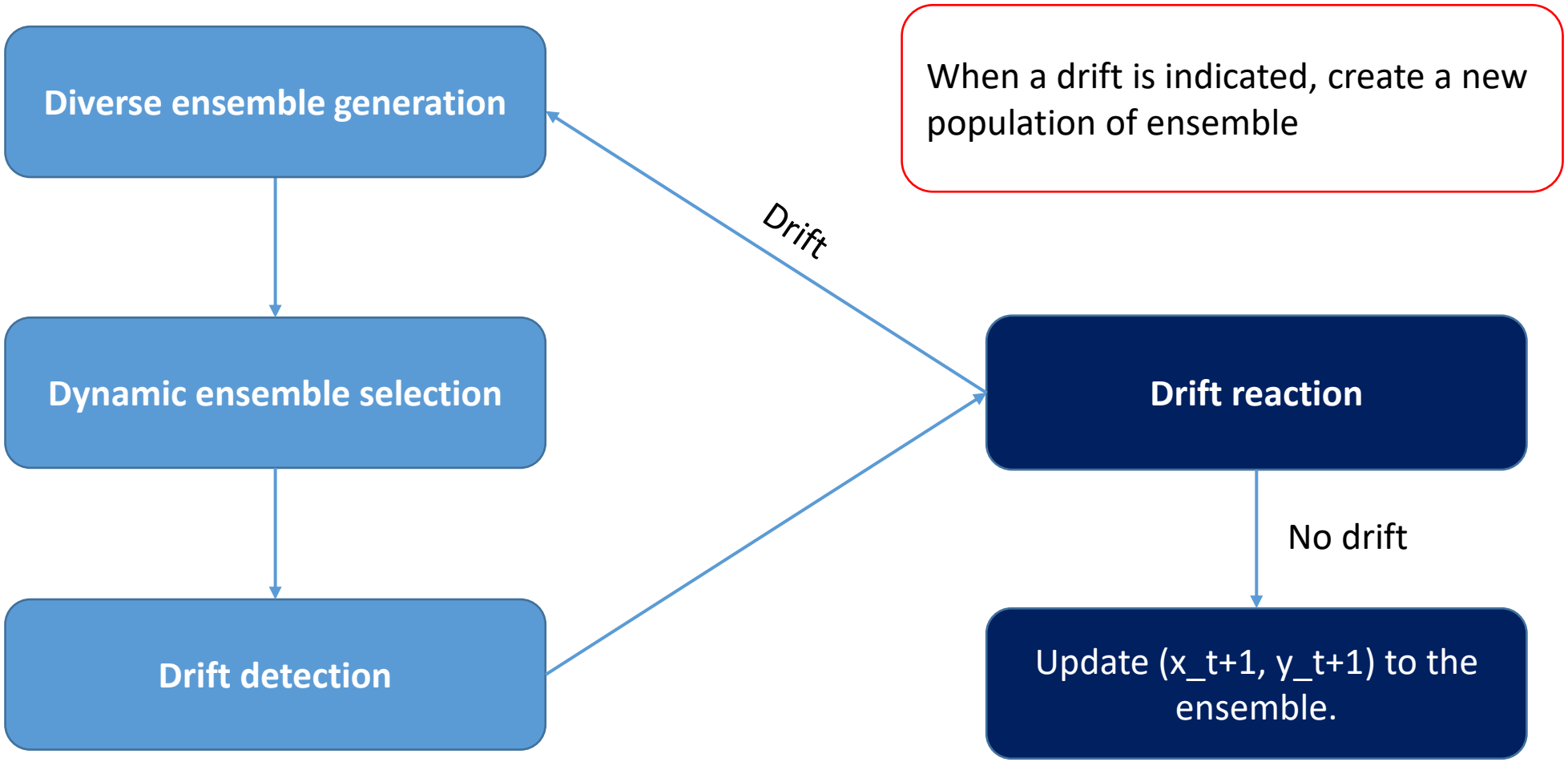
Method


Diverse ensemble generation
Dynamic ensemble selection
Drift detection
Drift reaction
Choose an expert ensemble with highest prequential (predictive and sequential) accuracy until instance (x_t) to assign a label (y_t+1) for unknown instance (x_t+1).

Method


Diverse ensemble generation
Dynamic ensemble selection
Drift detection
Drift reaction
Compare the true label and the predicted label to detect whether the prediction is correct or not, then through error monitoring process to indicate the occurrence of drift.

Method


Diverse ensemble generation
Dynamic ensemble selection
Drift detection
Drift
When a drift is indicated, create a new population of ensemble
Drift reaction
No drift
Update (x_t+1, y_t+1) to the ensemble.

Experiment

- To evaluate the accuracy performance of DESDD, three baseline are used.

1. DDM which employs a single classifier.
2. DDD which provide a population of ensembles.
3. Leveraging Bagging (LB), which generate an ensemble of classifier.

Dataset

- **Synthetic datasets**
  - Agrawal
  - SEA
- **Real-world datasets**
  - Forest Covertype
  - KDDCup
  - Poker-Hand
  - Spam

Synthetic datasets

Agrawal

- Create by Agrawal generator from python scikit-multiflow, which purposed by Agrawal ei al..
- Presumably these determine whether the loan should be approved.
- Include Abrupt and Gradual two datasets
- Represented by 9 features and 2 classes

Synthetic datasets

Agrawal Abrupt

	salary	commission	age	elevel	car	zipcode	hvalue	hyears	loan	target
0	20000.000000	10000.000000	25	3	13	0	0.0	13	117647.252277	1.0
1	123532.801577	0.000000	59	0	14	4	0.0	21	70157.813480	0.0
2	22283.573685	27839.376667	68	4	15	1	0.0	3	402571.281454	1.0
3	57610.885652	75000.000000	42	3	18	8	100000.0	18	0.000000	0.0
4	41205.891086	23198.635777	58	3	12	5	0.0	6	78517.837134	1.0
5	66216.155035	75000.000000	22	4	5	0	900000.0	9	155678.006962	0.0

- ✓ 10000 instances
- ✓ 9 features
- ✓ 2 classes

Agrawal Gradual

	salary	commission	age	elevel	car	zipcode	hvalue	hyears	loan	target
0	82491.139631	0.000000	49	2	5	2	0.0	30	367136.105915	1.0
1	33951.792240	58596.165885	33	1	1	0	0.0	11	397159.255491	0.0
2	104109.093591	0.000000	34	2	0	1	800000.0	5	0.000000	1.0
3	29530.160842	48373.632550	25	0	9	2	700000.0	3	291690.229487	0.0
4	116373.386172	0.000000	49	3	7	3	600000.0	11	23639.120009	1.0
5	41860.269091	50875.938550	29	1	0	2	0.0	16	301417.610963	0.0

- ✓ 10000 instances
- ✓ 9 features
- ✓ 2 classes

Synthetic datasets

SEA

- Create by SEA generator from python scikit-multiflow.
- Include Abrupt and Gradual two datasets
- Represented by 3 features and 3 classes

Synthetic datasets

SEA Abrupt

	0	1	2	3
0	3.750571	6.403046	9.500166	1.0
1	0.947599	3.946426	0.049449	0.0
2	9.558069	8.206094	3.449830	2.0
3	1.622260	2.900697	0.450082	2.0
4	7.365332	8.392115	7.093616	1.0
5	1.549115	6.325615	1.091434	0.0
6	4.961713	3.772385	9.790600	1.0
7	2.328748	4.919660	3.623373	0.0
8	7.251421	4.561816	0.607120	2.0
9	0.819152	2.656252	2.452066	0.0
10	6.745742	1.671433	4.781386	2.0

- ✓ 10000 instances
- ✓ 3 attributes
- ✓ 3 classes

SEA Gradual

	0	1	2	3
0	4.170220	7.203245	0.001144	1.0
1	1.467559	0.923386	1.862602	2.0
2	3.967675	5.388167	4.191945	2.0
3	1.403869	1.981015	8.007446	0.0
4	3.134242	6.923226	8.763892	1.0
5	0.850442	0.390548	1.698304	0.0
6	5.331653	6.918771	3.155156	1.0
7	1.300286	0.193670	6.788355	2.0
8	6.997584	1.023344	4.140560	1.0
9	4.141793	0.499535	5.358964	0.0
10	5.148891	9.445948	5.865550	2.0

- ✓ 10000 instances
- ✓ 3 features
- ✓ 3 classes

Real-world datasets

Forest Covertype

- Contains tree observations from four areas of the Roosevelt National Forest in Colorado.
- Represented by 54 features and 7 classes
- Collected from http://archive.ics.uci.edu/ml//datasets/Covertype

Real-world datasets

Forest Covertype

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways	Hillshade_9am	Hillshade_N
0	2596	51	3	258	0	510	221	
1	2590	56	2	212	-6	390	220	
2	2804	139	9	268	65	3180	234	
3	2785	155	18	242	118	3090	238	
4	2595	45	2	153	-1	391	220	
5	2579	132	6	300	-15	67	230	
6	2606	45	7	270	5	633	222	
7	2605	49	4	234	7	573	222	
8	2617	45	9	240	56	666	223	
9	2612	59	10	247	11	636	228	
10	2612	201	4	180	51	735	218	
11	2886	151	11	371	26	5253	234	
12	2742	134	22	150	69	3215	248	
13	2609	214	7	150	46	771	213	
14	2503	157	4	67	4	674	224	
15	2495	51	7	42	2	752	224	

- ✓ 581,012 instances
- ✓ 54 features
- ✓ 7 classes

Real-world datasets

KDD Cup 1999

- Represent several network intrusion problems simulated in a military network.
- Represented by 41 features and 2 classes
- Collected from https://datahub.io/machine-learning/kddcup99

Real-world datasets

KDD Cup 1999

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	...	dst_host_srv_count	dst_host_same_srv_rate	dst_host_
0	0	tcp	http	SF	181	5450	0	0	0	0	...	9	1.0	
1	0	tcp	http	SF	239	486	0	0	0	0	...	19	1.0	
2	0	tcp	http	SF	235	1337	0	0	0	0	...	29	1.0	
3	0	tcp	http	SF	219	1337	0	0	0	0	...	39	1.0	
4	0	tcp	http	SF	217	2032	0	0	0	0	...	49	1.0	
5	0	tcp	http	SF	217	2032	0	0	0	0	...	59	1.0	
6	0	tcp	http	SF	212	1940	0	0	0	0	...	69	1.0	
7	0	tcp	http	SF	159	4087	0	0	0	0	...	79	1.0	
8	0	tcp	http	SF	210	151	0	0	0	0	...	89	1.0	
9	0	tcp	http	SF	212	786	0	0	0	1	...	99	1.0	
10	0	tcp	http	SF	210	624	0	0	0	0	...	109	1.0	
11	0	tcp	http	SF	177	1985	0	0	0	0	...	119	1.0	
12	0	tcp	http	SF	222	773	0	0	0	0	...	129	1.0	
13	0	tcp	http	SF	256	1169	0	0	0	0	...	139	1.0	
14	0	tcp	http	SF	241	259	0	0	0	0	...	149	1.0	
15	0	tcp	http	SF	260	1837	0	0	0	0	...	159	1.0	

- ✓ 494,020 instances
- ✓ 41 features
- ✓ 2 classes

Real-world datasets

Poker-Hand

- Each record is an example of a hand consisting of five playing cards drawn from a standard deck of 52.
- Each card is described using two attributes (suit and rank), for a total of 10 predictive attributes
- Represented by 11 features and 10 classes
- Collected from https://datahub.io/machine-learning/kddcup99

Real-world datasets

Poker-Hand

	S1	C1	S2	C2	S3	C3	S4	C4	S5	C5	class
0	1	1	1	13	2	4	2	3	1	12	0
1	3	12	3	2	3	11	4	5	2	5	1
2	1	9	4	6	1	4	3	2	3	9	1
3	1	4	3	13	2	13	2	1	3	6	1
4	3	10	2	7	1	2	2	11	4	9	0
5	1	3	4	5	3	4	1	12	4	6	0
6	2	6	4	11	2	3	4	9	1	7	0
7	3	2	4	9	3	7	4	3	4	5	0
8	4	4	3	13	1	8	3	9	3	10	0
9	1	9	3	8	4	4	1	7	3	5	0
10	4	7	3	12	1	13	1	9	2	6	0
11	2	12	1	3	2	11	2	7	4	8	0
12	4	2	2	9	2	7	1	5	3	11	0
13	1	13	2	6	1	6	2	11	3	5	1
14	3	8	2	7	1	9	3	6	2	3	0
15	2	10	1	11	1	9	3	1	1	13	0

- ✓ 1,000,000 instances
- ✓ 11 features
- ✓ 10 classes

Real-world datasets

SPAM

- This SPAM data is about E-mail.
- Represented by 57 features and 2 classes
- Collected from https://archive.ics.uci.edu/ml/datasets/spambase

Real-world datasets

SPAM

	word_freq_make	word_freq_address	word_freq_all	word_freq_3d	word_freq_our	word_freq_over	word_freq_remove	word_freq_internet	word_freq_orde
0	0.00	0.64	0.64	0.0	0.32	0.00	0.00	0.00	0.0(
1	0.21	0.28	0.50	0.0	0.14	0.28	0.21	0.07	0.0(
2	0.06	0.00	0.71	0.0	1.23	0.19	0.19	0.12	0.6
3	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.3
4	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.3
5	0.00	0.00	0.00	0.0	1.85	0.00	0.00	1.85	0.0(
6	0.00	0.00	0.00	0.0	1.92	0.00	0.00	0.00	0.0(
7	0.00	0.00	0.00	0.0	1.88	0.00	0.00	1.88	0.0(
8	0.15	0.00	0.46	0.0	0.61	0.00	0.30	0.00	0.9
9	0.06	0.12	0.77	0.0	0.19	0.32	0.38	0.00	0.0(
10	0.00	0.00	0.00	0.0	0.00	0.00	0.96	0.00	0.0(
11	0.00	0.00	0.25	0.0	0.38	0.25	0.25	0.00	0.0(
12	0.00	0.69	0.34	0.0	0.34	0.00	0.00	0.00	0.0(
13	0.00	0.00	0.00	0.0	0.90	0.00	0.90	0.00	0.0(
14	0.00	0.00	1.42	0.0	0.71	0.35	0.00	0.35	0.0(
15	0.00	0.42	0.42	0.0	1.27	0.00	0.42	0.00	0.0(

- ✓ 4601 instances
- ✓ 57 features
- ✓ 2 classes

Thank you for your attention.