HDWM-Experiment with Synthetic Drift Streams

Using Clustering in HDWM Algorithm

UNIVERSITY OF READING

Supervisors: Frederic Stahl, Atta Baadi

PhD Student: Mobin Idrees

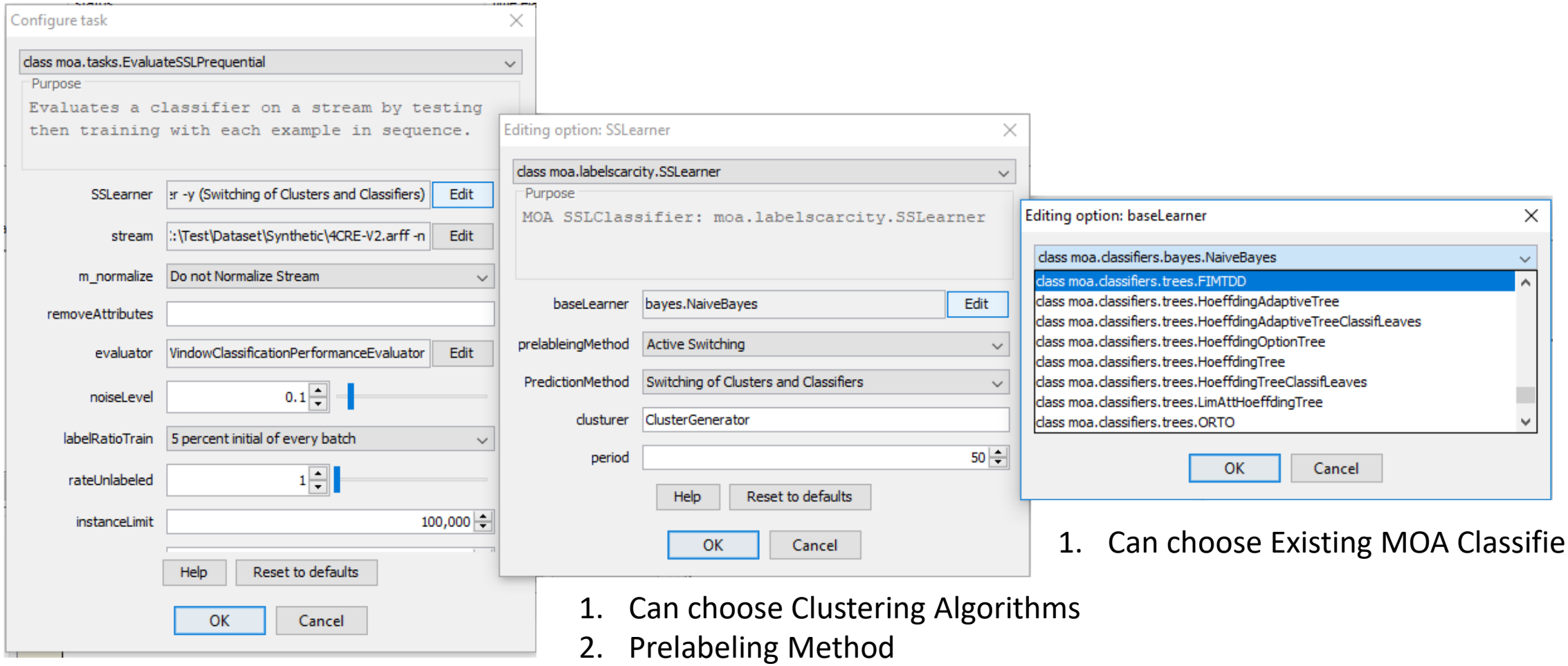
Recent Work

Part 1: Modification in Existing HDWM algorithm

- **Included ‘Adaptive Random Forest’ algoritm in the experiments.**
- **In previous HDWM paper, the approach was not able to improve the predictive performance in some datasets.**
- **HDWM algorithm is update to improve the predictive performance (%) in less time (CPU seconds)**

Part 2: Experiments on Label Scarcity on Modified HDWM algorithm

SemiSupervised Prequential Task



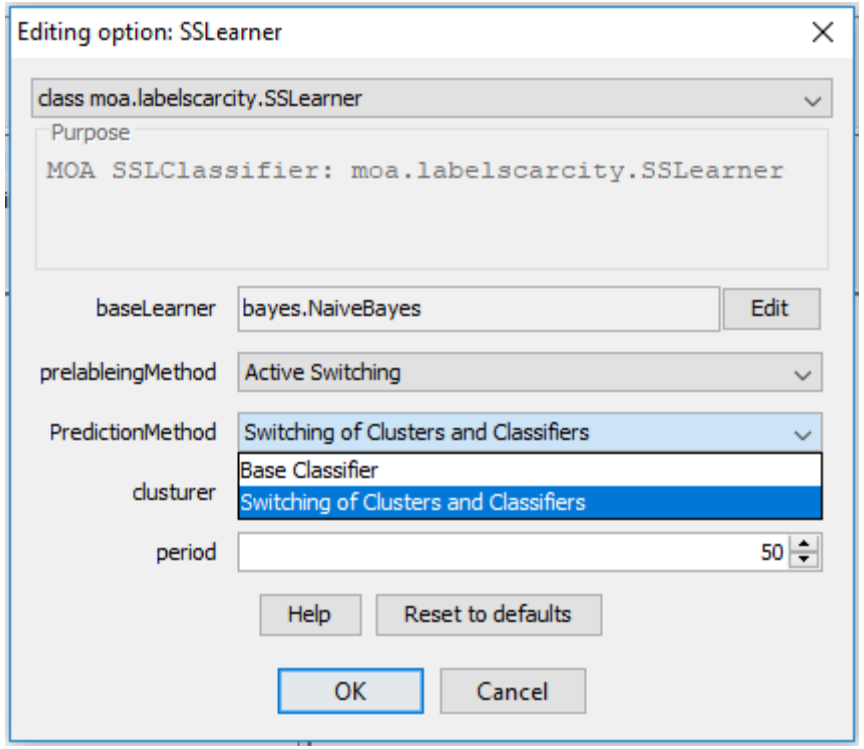
1. Can choose Existing MOA Classifiers

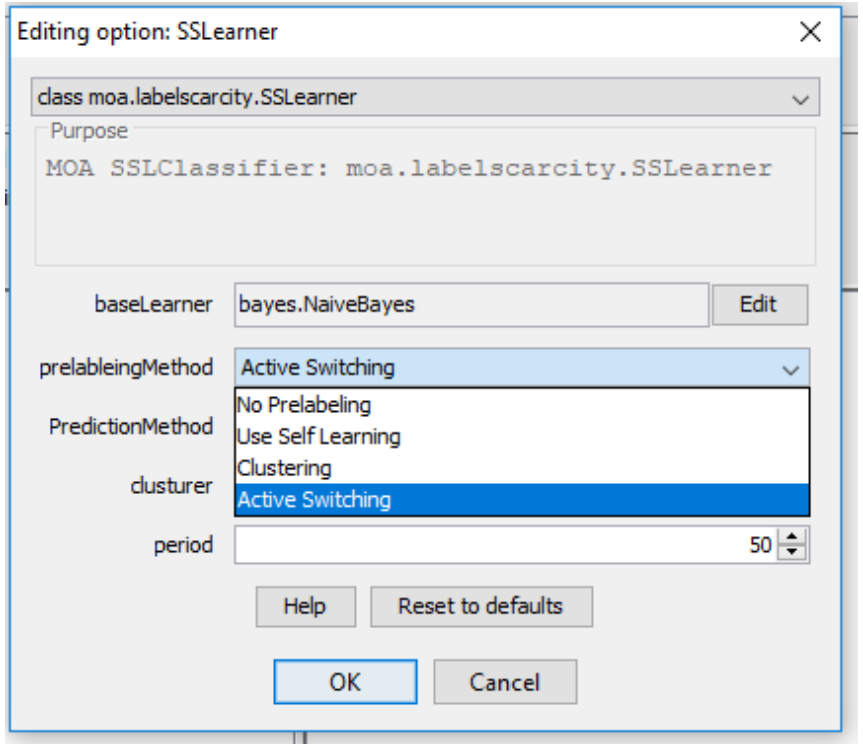
1. Can choose Clustering Algorithms
2. Prelabeling Method
3. Final Prediction Method

In the Previous version we had to develop Semisupervised classifiers using the existing Supervervised Classsifiers in MOA.
For instance, we had to implement SSL_NaiveBase for NaiveBayes and likewise SSL_HoeffdingTree for HoeffdingTree and so on.

By Creating a new SSLerner Class, we can now choose any combination of Supervised and clustering algorithms.

Two New Parameters





Parameter - x `Prelabeling Stratedy {"No Prelabeling", "Use Self Learning", "Clustering", "Active Switching"}`

Parameter -y `Final Prediction { “Base Classifier", "Switching of Clusters and Classifiers"}`

Switching Mechanism (First Batch 100% Labeled, followed by 5% labeled instances in each batch

Swithing in Prelabeling and No Switching in Final Vote (Classifiers Only)

Prelabeling `-x {"Active Switching"}`

Final Vote `-y {"Base Classifier"}`

	Naïve Bayes		
	NO prelabeling	Prelabeled	Gain/Loss
UG_2C_2D	52.516	68.357	+15.841
UG_2C_3D	62.821	68.286	+5.465
1CSurr	63.943	73.329	+9.386
4CR	19.771	34.076	+14.305
4CRE-V2	24.080	32.892	+8.812
stagger	52.808	52.925	+0.117
RandomTree	62.783	62.873	+0.090
LED	44.387	48.655	+4.268
Hyperplane	61.813	70.045	+8.232
SEA_Mixed	82.408	84.972	+2.564
RandomRBF	25.666	39.25	+13.584
AGRAWAL_s	64.598	66.492	+1.894
AGRAWAL_Mixed	65.435	67.422	+1.987
sensor	14.878	18.263	+3.385
covtype	63.19	70.50	+7.306

Swithing in Prelabeling and Switching in Final Vote (Cluster or Classifier)

Prelabeling `-x {"Active Switching"}`

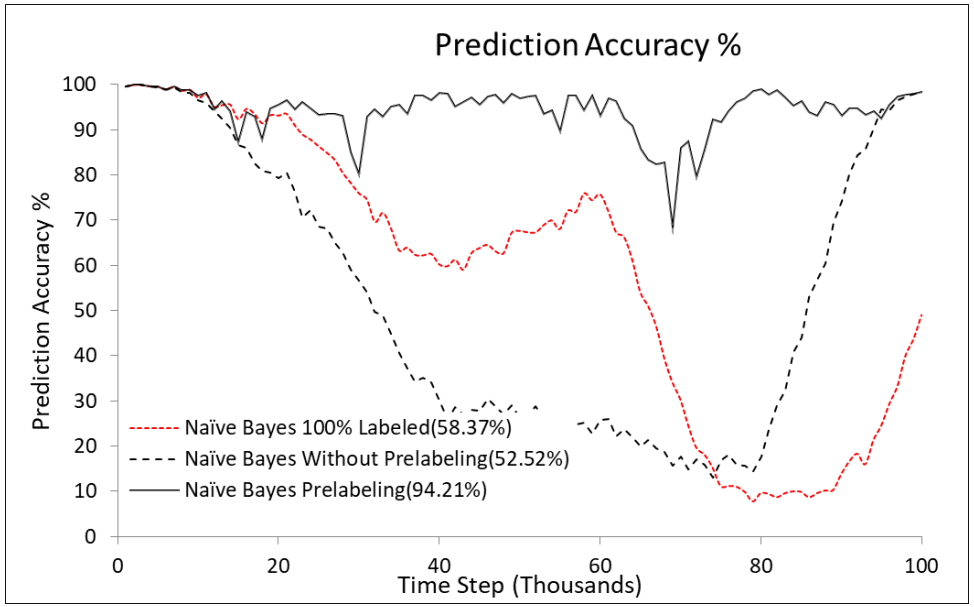
Final Vote `-y {"Switching of Clusters and Classifiers"}`

	Naïve Bayes		
	NO prelabeling	Prelabeled	Gain/Loss
UG_2C_2D	52.516	94.209	+41.693
UG_2C_3D	62.821	92.663	+29.842
1CSurr	63.943	85.971	+22.029
4CR	19.771	95.763	+75.992
4CRE-V2	24.080	84.766	+60.686
stagger	52.808	55.200	+2.392
RandomTree	62.783	66.109	+3.326
LED	44.387	48.678	+4.291
Hyperplane	61.813	68.942	+7.129
SEA_Mixed	82.408	82.59	+0.182
RandomRBF	25.666	74.764	+49.098
AGRAWAL_s	64.598	67.62	+3.022
AGRAWAL_Mixed	65.435	67.518	+2.083
sensor	14.878	22.123	+7.245
covtype	63.19	69.47	+6.280

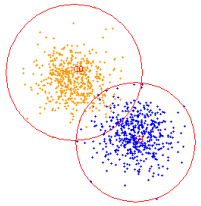
Switching Mechanism for Prelabeling (Naïve Bayes)

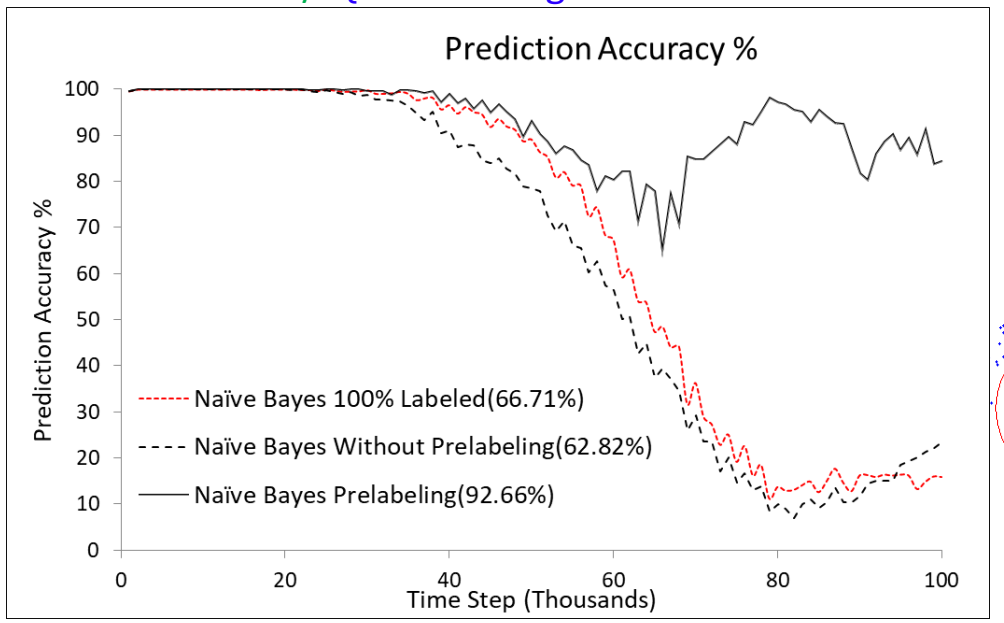
Prelabeling – x {"Active Switching"}

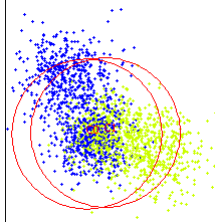
Final Vote – y {"Switching of Clusters and Classifiers"}

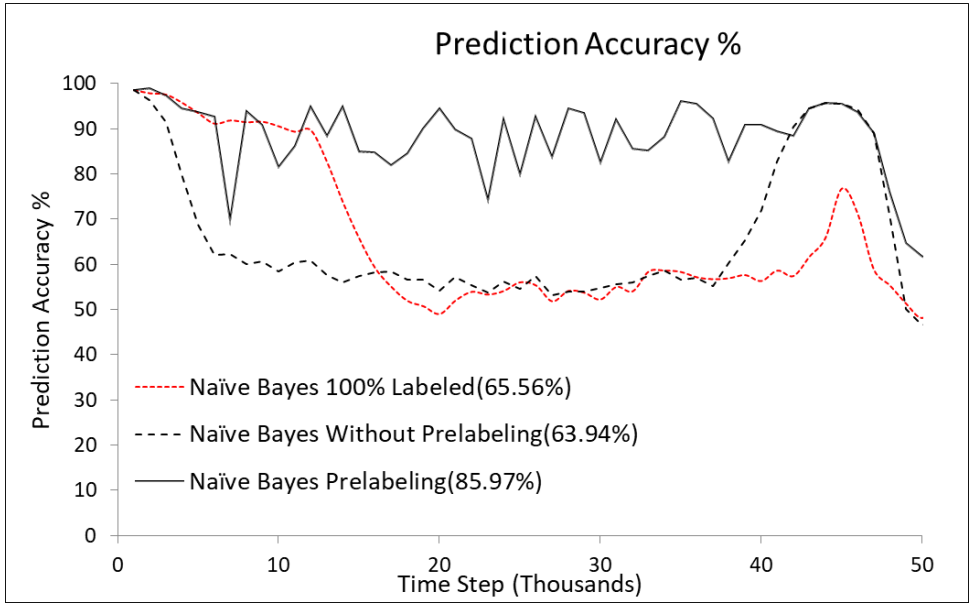


UG_2C_2D

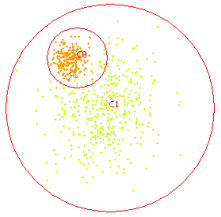


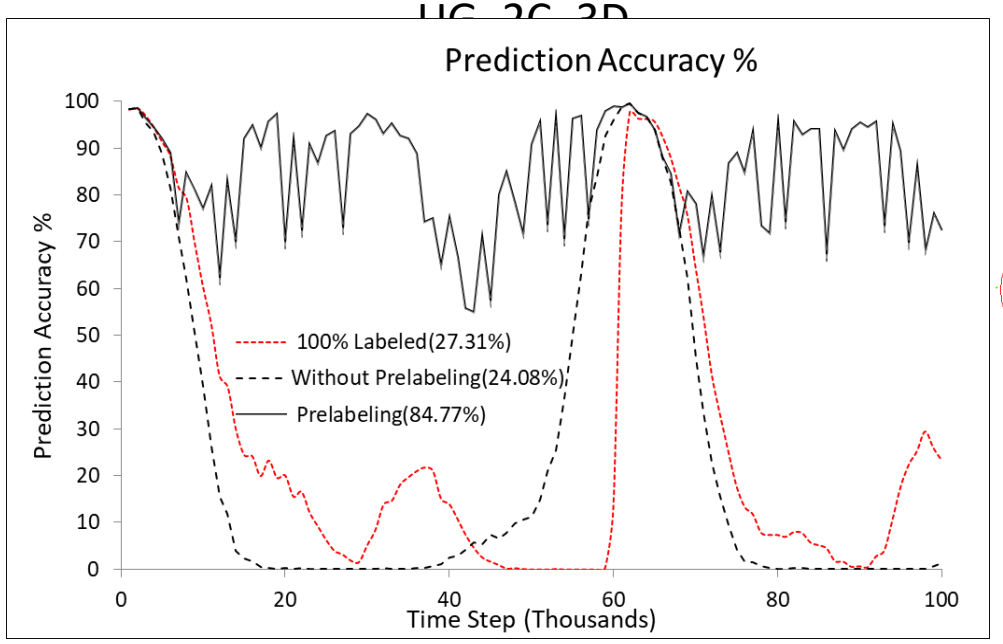


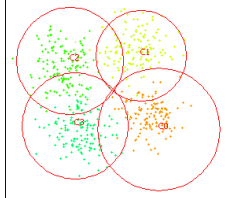




1CSurr





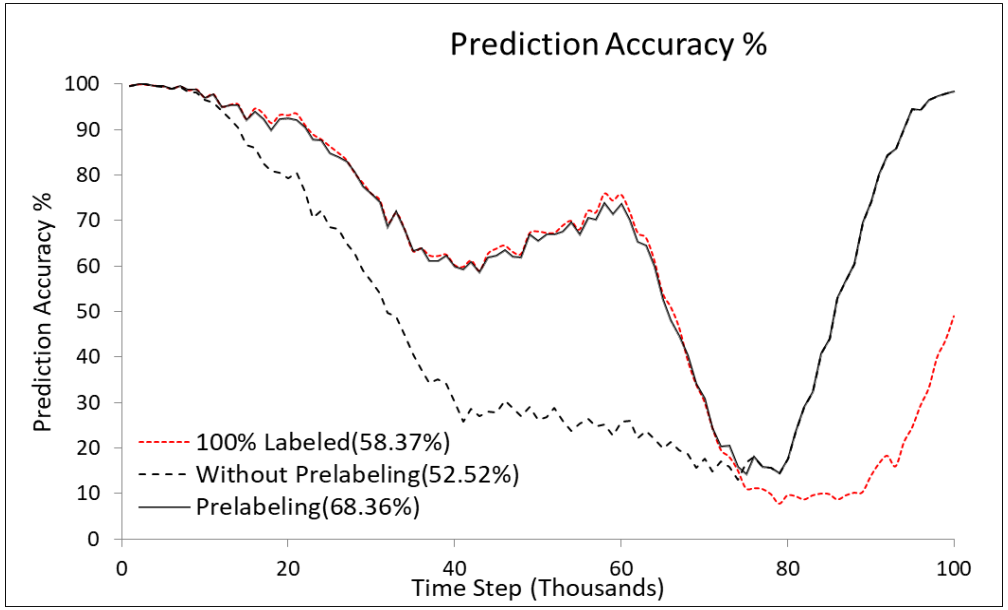


4CRE-V2

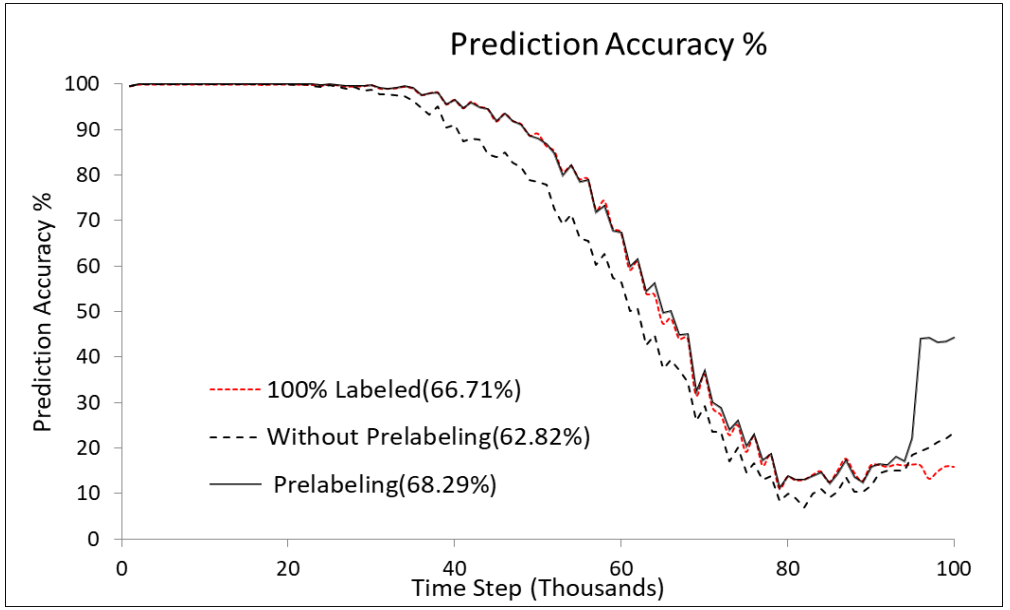
Switching Mechanism for Prelabeling (Naïve Bayes)

Prelabeling – x {"Active Switching"}

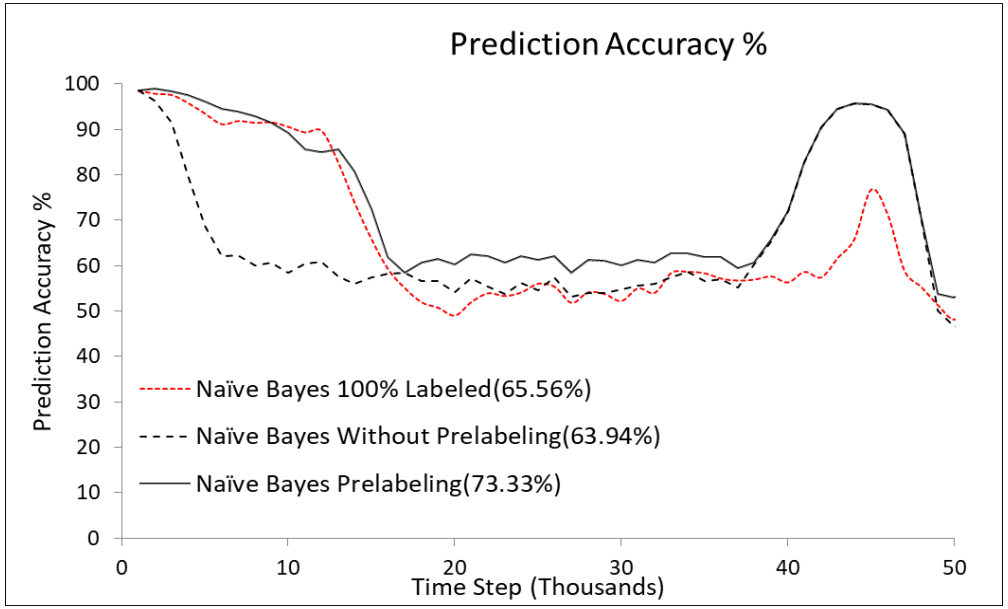
Final Vote – y {"Base Classifier"}



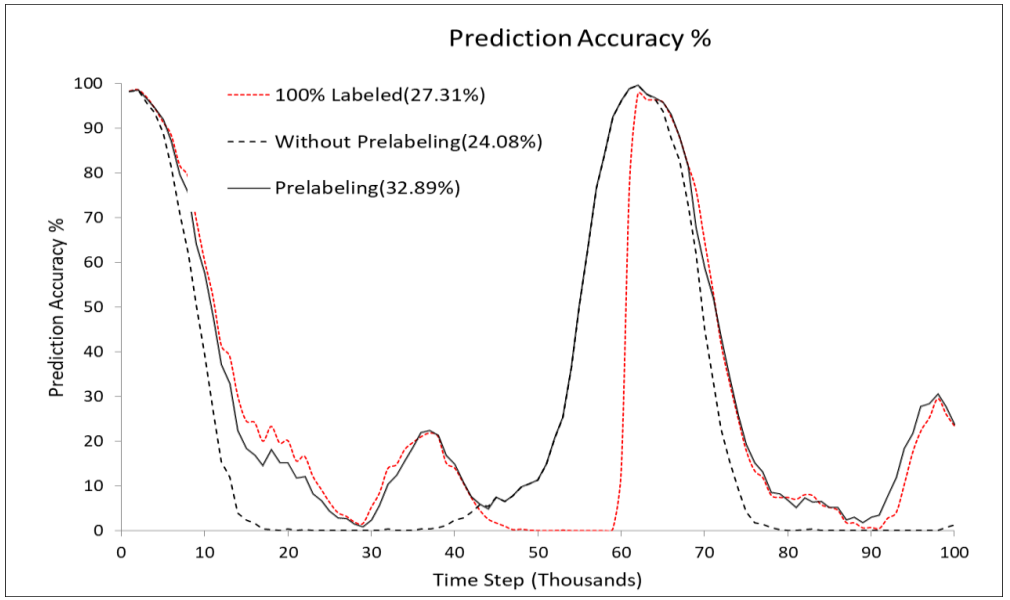
UG_2C_2D



UG_2C_3D

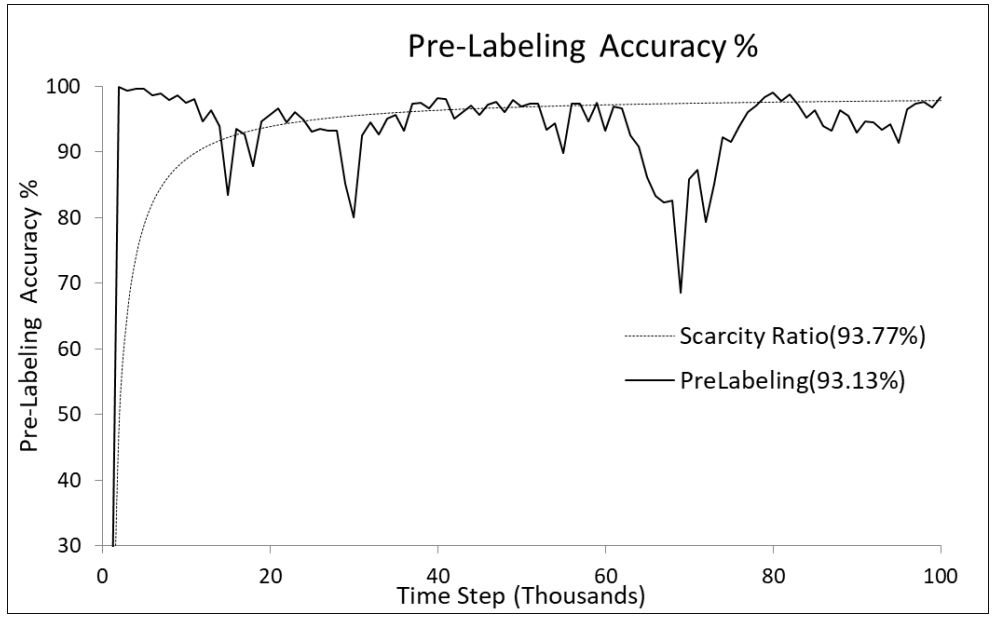


1CSurr

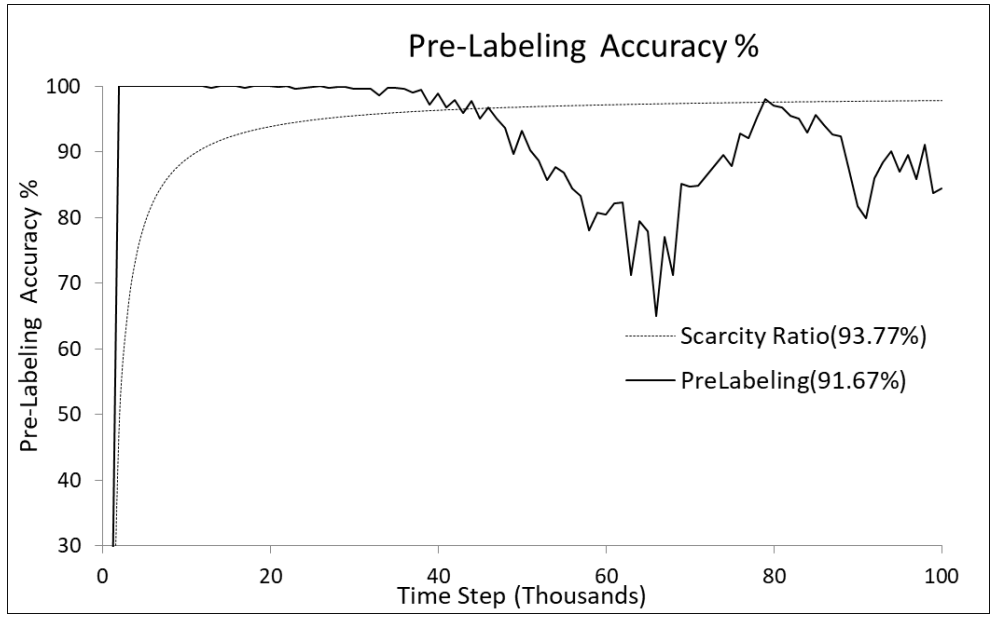


4CRE-V2

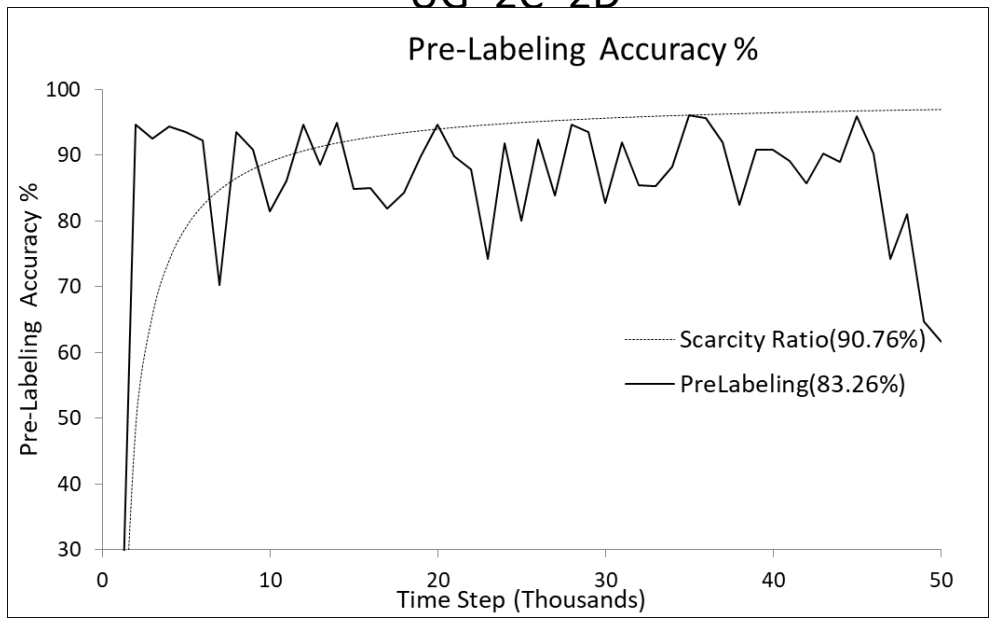
Prelabeling ACC(%) (Naïve Bayes)



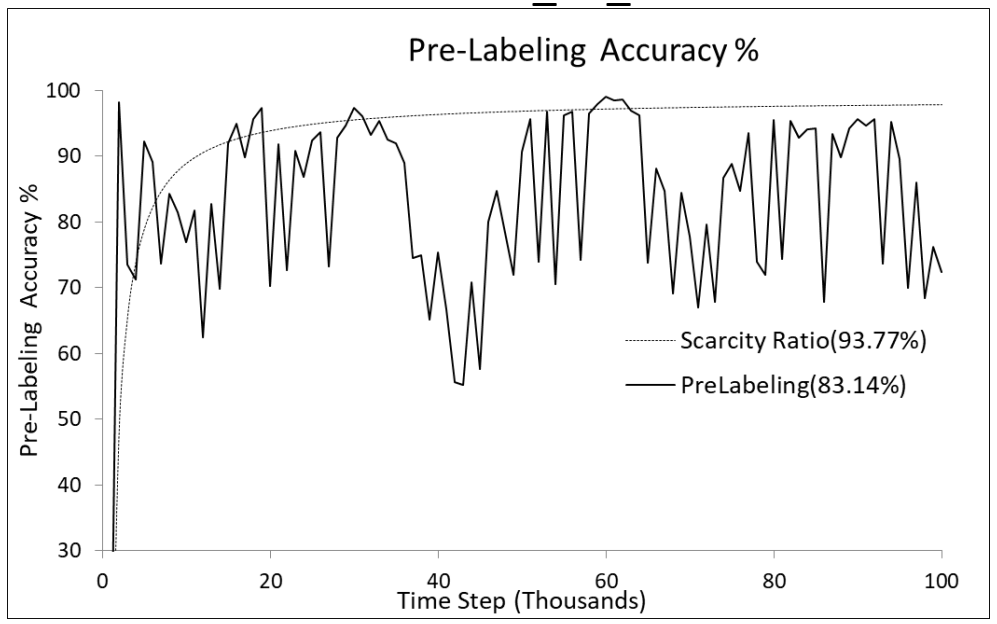
UG_2C_2D



UG_2C_3D



1CSurr



4CRE-V2

MOA Graphical User Interface

Label Scarcity | Classification | Clustering | Active Learning

Configure | moa.tasks.EvaluateSSLPrequential -l (moa.labelscarcity.SSLearner -y (Switching of Clusters and Classifiers)) -s (clustering.FileStream -f C:\Test\Dataset\Synthetic\4CRE-V2.arff -n) -i 100000 -f 1000 | Run Now

command	status	time elapsed	current activity	% complete
EvaluateSSLPrequential	paused	6.31s	Evaluating learner...	97.88
EvaluateSSLPrequential	completed	6.86s		100.00
EvaluateSSLPrequential	completed	10.44s		100.00

Pause | Resume | Cancel | Delete

Preview (6.28s) | Refresh | Auto refresh: every second

```
learning evaluation instances,Scarcity Ratio,evaluation time (cpu seconds),model cost (RAM-Hours),classified instanc
1000.0,0.0,0.03125,9.926346441109975E-9,1000.0,98.3,97.73333333333333,97.8120978120978,97.6356050069541,1000.0,12278
2000.0,49.45,0.078125,2.7349645582338175E-8,2000.0,98.6,98.13333333333333,98.14814814814815,97.98561151079137,2000.0
3000.0,65.93333333333334,0.140625,5.0580711103975774E-8,3000.0,96.6,95.46666666666667,95.49071618037135,95.423956931
4000.0,74.175,0.203125,7.38257464642326E-8,4000.0,94.5,92.66666666666666,92.69588313413014,92.50681198910081,4000.0,
5000.0,79.12,0.25,9.125952298442522E-8,5000.0,92.10000000000001,89.46666666666667,89.36742934051145,89.0429958391123
6000.0,82.41666666666667,0.328125,1.2031581718474625E-7,6000.0,89.0,85.33333333333334,84.82758620689656,84.827586206
7000.0,84.77142857142857,0.40625,1.493721113850673E-7,7000.0,73.8,65.06666666666666,64.6900269541779,64.109589041095
8000.0,86.5375,0.46875,1.7261714674532413E-7,8000.0,84.7,79.6,80.40973111395647,79.01234567901234,8000.0,1437648.0,1
```

Export as .txt file...

Evaluation

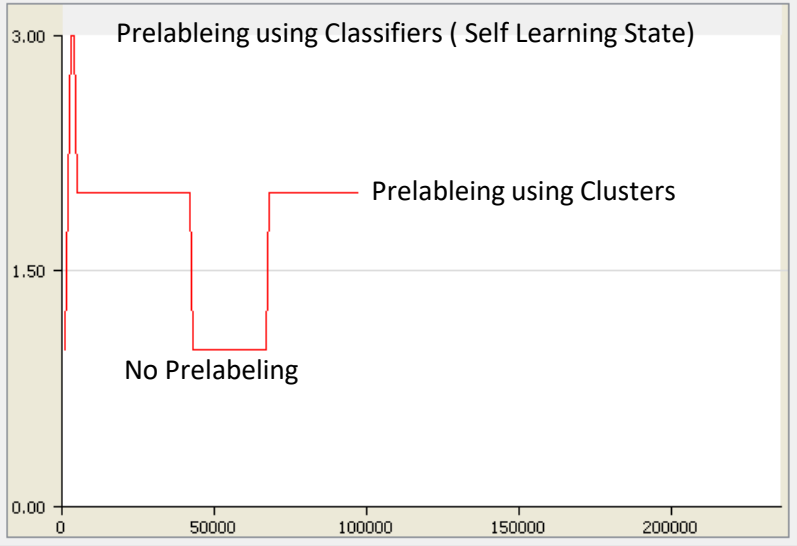
Values

Measure	Current		Mean	
Accuracy	86.10	-	85.15	-
Scarcity	97.88	-	93.64	-
Pre-Labeling Ratio	86.00	-	83.47	-
Cluster Size	3.00	-	3.77	-
Ensemble Size	1.00	-	1.00	-
Labeling State	2.00	-	1.75	-
Ram-Hours	6.14	-	3.03	-
Time	97.88	-	93.64	-
Memory	0.09	-	0.05	-

Plot

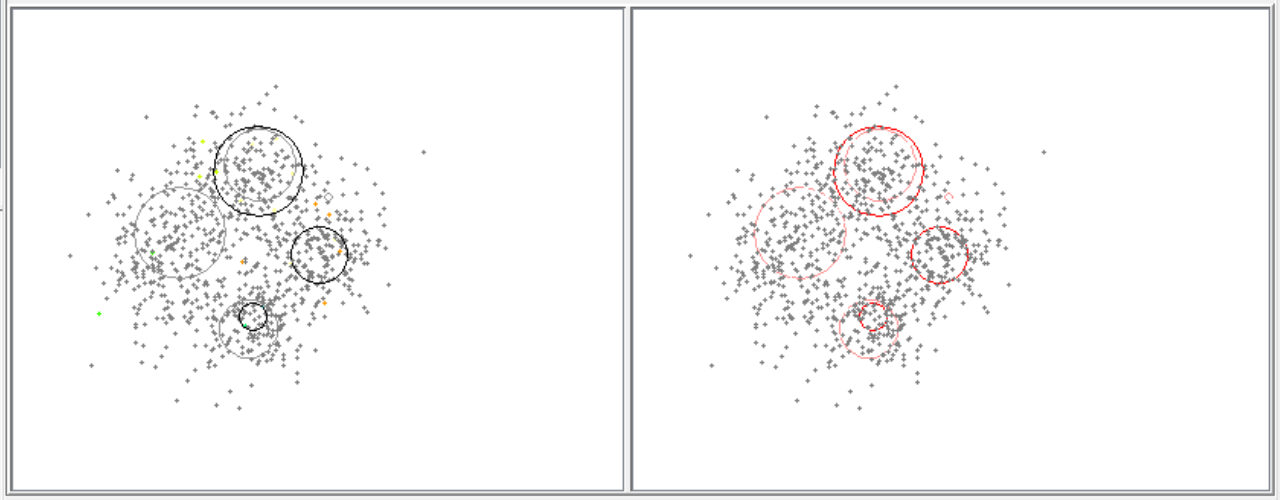
Zoom in Y | Zoom out Y | Zoom in X | Zoom out X

PreLabeling State



Visualization | Evaluation Setup

X Dim 1 | Y Dim 2 | Points | Microclustering | Ground truth | Clustering | Visualisation Speed | Processed:97800 | Pause in: 5000 | Cluster: ClusterGenerator



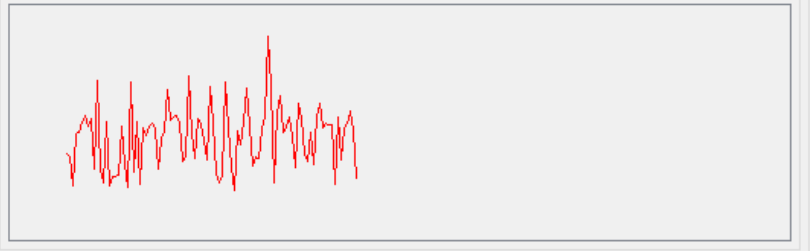
Evaluation

Values

Measure	Current		Mean	
F1-P	0.36	-	0.36	-
F1-R	0.33	-	0.26	-
Purity	0.95	-	0.96	-
GPrecision	1.00	-	1.00	-
GRecall	0.26	-	0.37	-
Redundancy	0.00	-	0.01	-
numCluster	3.00	-	3.78	-
numClasses	3.00	-	3.78	-

Plot

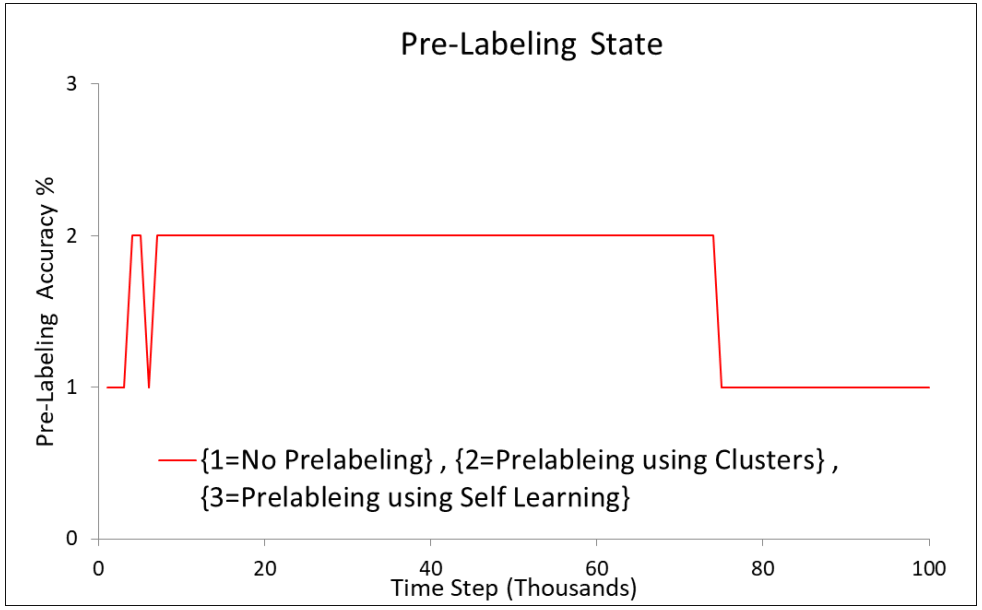
Zoom in Y | Zoom out Y | Zoom in X | Zoom out X



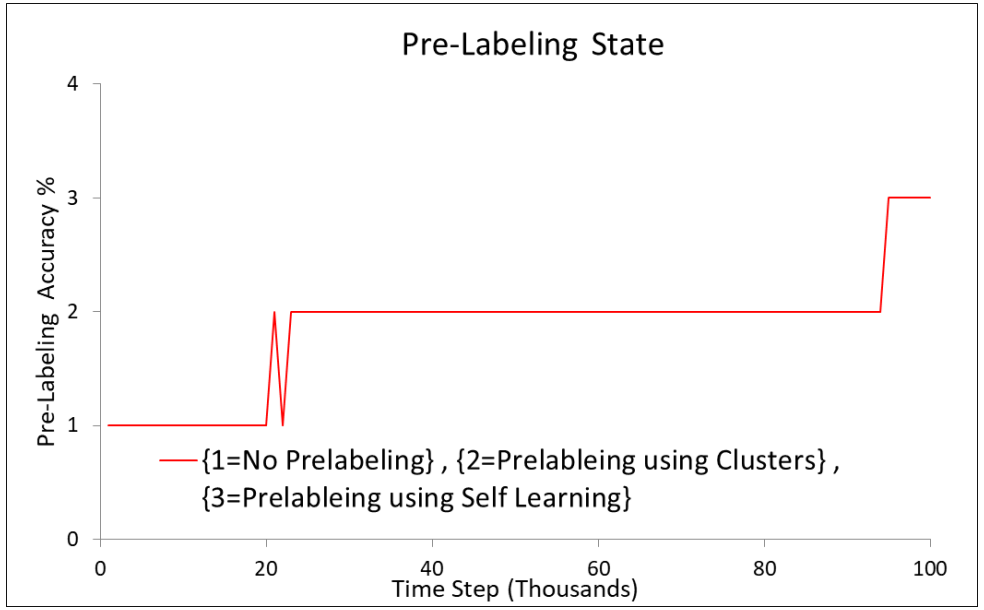
Switching States for Prelabeling (Naïve Bayes)

Prelabeling – x {"Active Switching"}

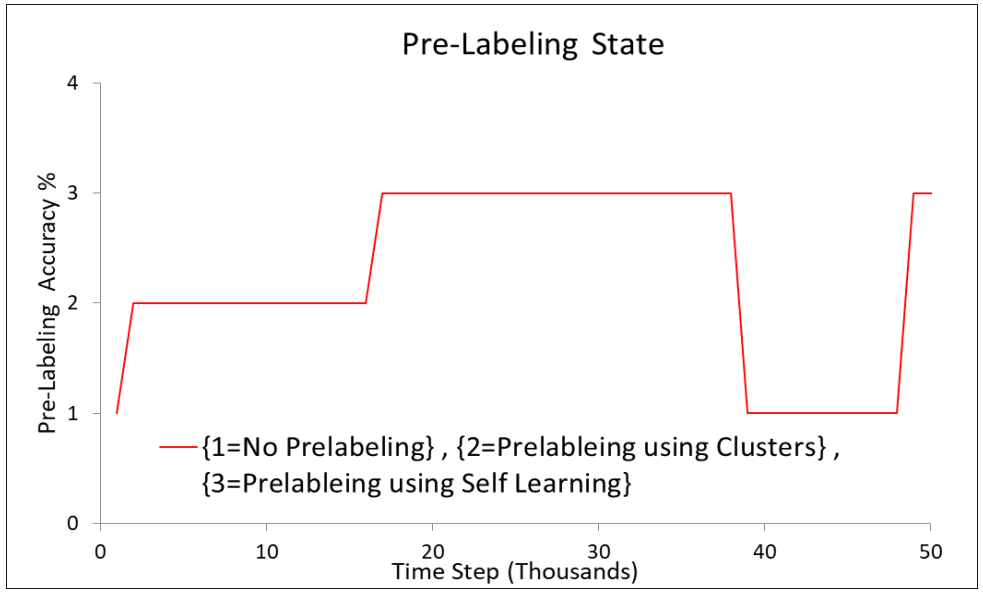
Final Vote – y {"Base Classifier"}



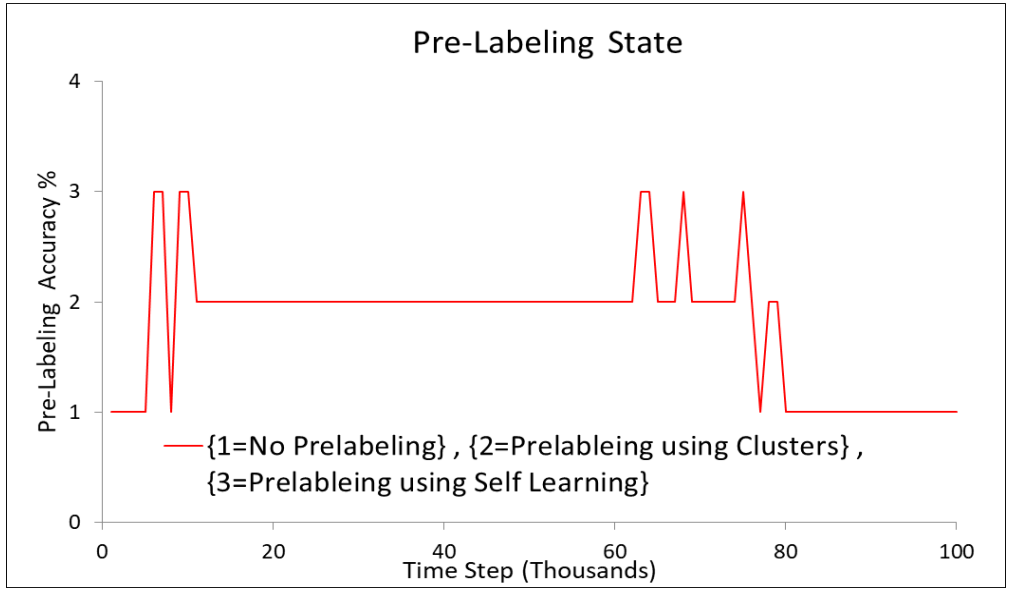
UG_2C_2D



UG_2C_3D



1CSurr



4CR