MineClass: A Synergy of Data Stream Classification and Novel Class Detection

Mohammad Mehedy Masud†, Jing Gao‡, Latifur Khan†, Jiawei Han‡ and Bhavani Thuraisingham†

†University of Texas at Dallas, ‡University of Illinois at Urbana Champaign

Introduction

Data Stream Classification faces three major problems:

- **Infinite length:** solution – incremental learning
- **Concept-drift:** solution – adapting to the most recent concept
- **Novel class:** solution - *MineClass*

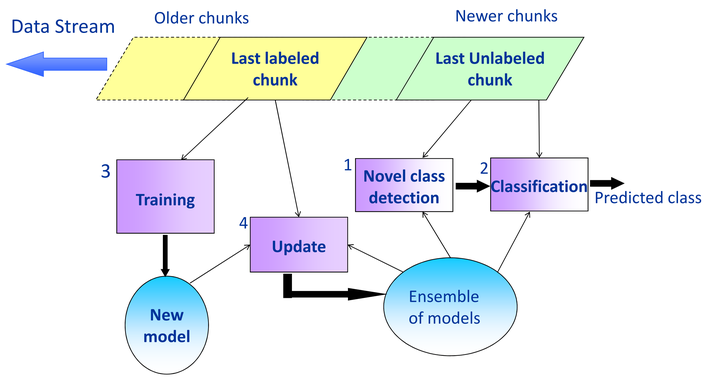


Figure 1. Overview of MineClass

Applications of novel class detection:

- Intrusion detection: detecting new kind of attack
- Text classification: detecting new category of text
- Fault detection: detecting new kind of fault

Basic idea

Assumption: A data point should be closer to the data points of its own class (cohesion) and farther apart from the data points of any other classes (separation).

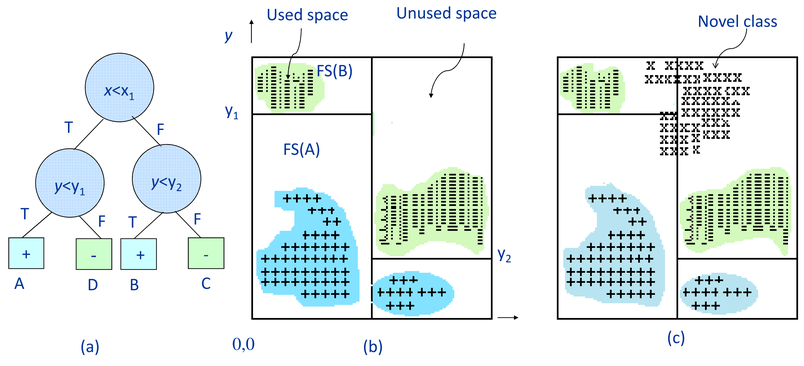


Figure 2. A decision tree (a), corresponding feature-space partitioning (b), and arrival of a novel class in the unused portions of the feature space (c).

Novel class detection

Steps:

1. **Save inventory of used spaces during training**
2. **Outlier detection and filtering**
3. **Measuring cohesion and separation**

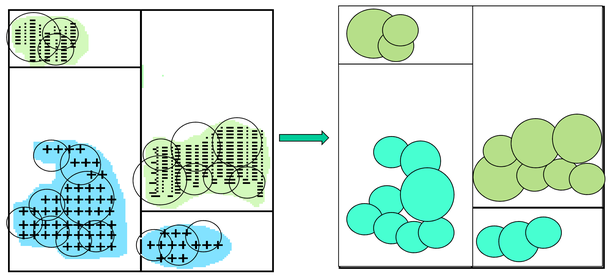


Figure 3. Saving inventory of used spaces as micro-clusters

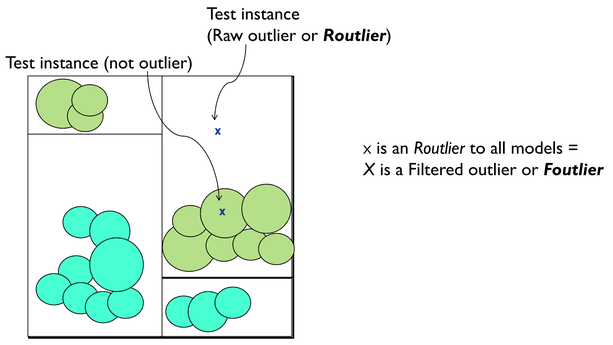
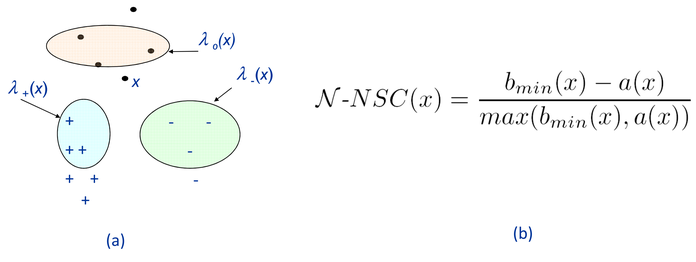


Figure 4. Outlier detection and filtering



$$\mathcal{N}\text{-}NSC(x) = \frac{b_{min}(x) - a(x)}{max(b_{min}(x), a(x))}$$

Figure 5. Measuring cohesion and separation by computing (a) λ_c neighborhood for *N*=3, and (b) *N*-neighborhood silhouette coefficient (*N*-NSC). A novel class is declared if *N*-NSC(.) is positive for at least *N* instances

Experiments and results

We evaluated our approach on two synthetic and two real datasets

Baseline: MineClass (MC), WCE-OLINDDA_PARALLEL(W-OP): WCE-OLINDDA_SINGLE(W-OS)

WCE-OLINDDA is a combination of the Weighted Classifier Ensemble (WCE) [1] and novel class detector OLINDDA [2]

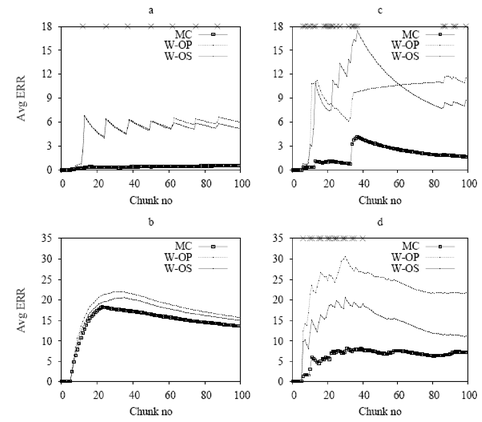


Figure 5. Error comparison on (a) SynCN, (b) SynC, (c) KDD and (d) Forest Cover.

Table-I. Run-time comparison on all datasets

Dataset	Time(sec)/chunk			Points/sec			Speed gain	
	MC	W-OP	W-OS	MC	W-OP	W-OS	MC over W-OP	MC over W-OS
SynC	**0.18**	0.81	0.19	**5,446**	1,227	5,102	**4**	**1**
SynCN	**0.27**	52.9	7.34	**3,656**	18	135	**203**	**27**
KDD	**0.95**	1369.5	222.8	**4,190**	2	17	**2,095**	**246**
Forest cover	**2.11**	213.1	10.79	**1,899**	18	370	**105**	**5**

Conclusion

MineClass handles all three problems in data stream classification:

- Handles infinite memory and concept-drift problem using ensemble classification
- Novel class detection problem using "class mining"
- It is an unique combination of stream classification and novel class detection

References

[1] E. J. Spinosa, A. P. de Leon F. de Carvalho, and J. Gama. Olindda: a cluster-based approach for detecting novelty and concept drift in data streams. In Proc. 2007 ACM symposium on Applied computing, pages 448–452, 2007.

[2] H. Wang, W. Fan, P. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In Proc. SIGKDD, 2003.