Partial Undersampling of Imbalanced Data for Cyber Threats Detection

Presented by:
Md Moniruzzaman

PhD Student, ICSL

Federation University Australia

Outline

- Introduction
- Existing method
- Proposed method
- Experimental results
- Conclusion and Future Works

Introduction

- Imbalanced dataset: A dataset can be categorized as imbalanced when the number of instances of some classes are significantly larger compared to the number of instances of other classes.

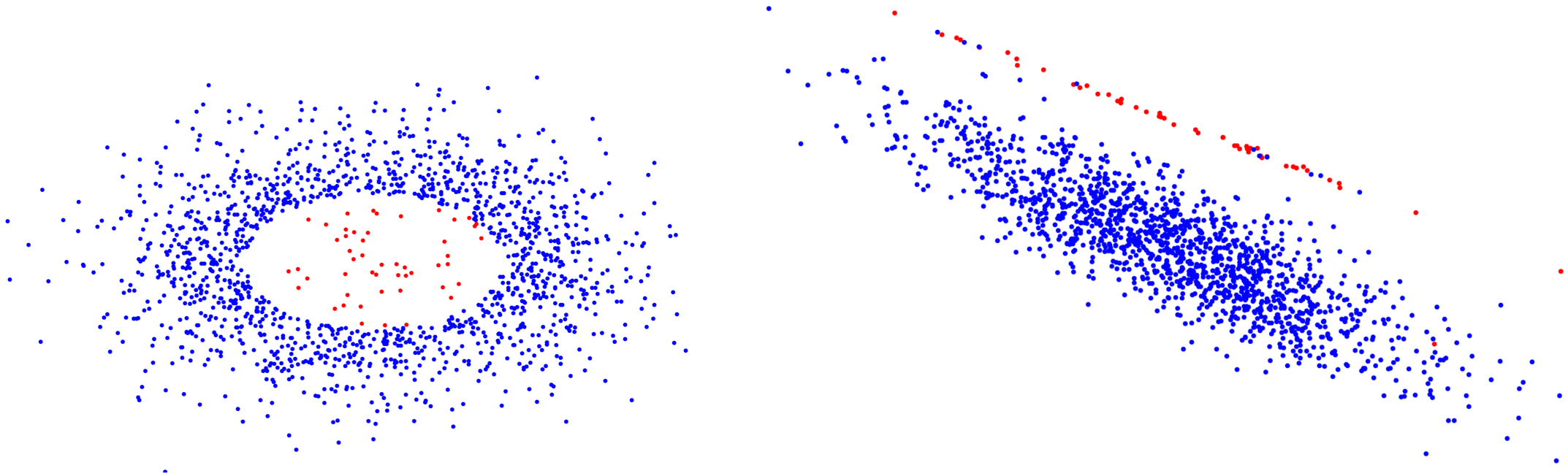
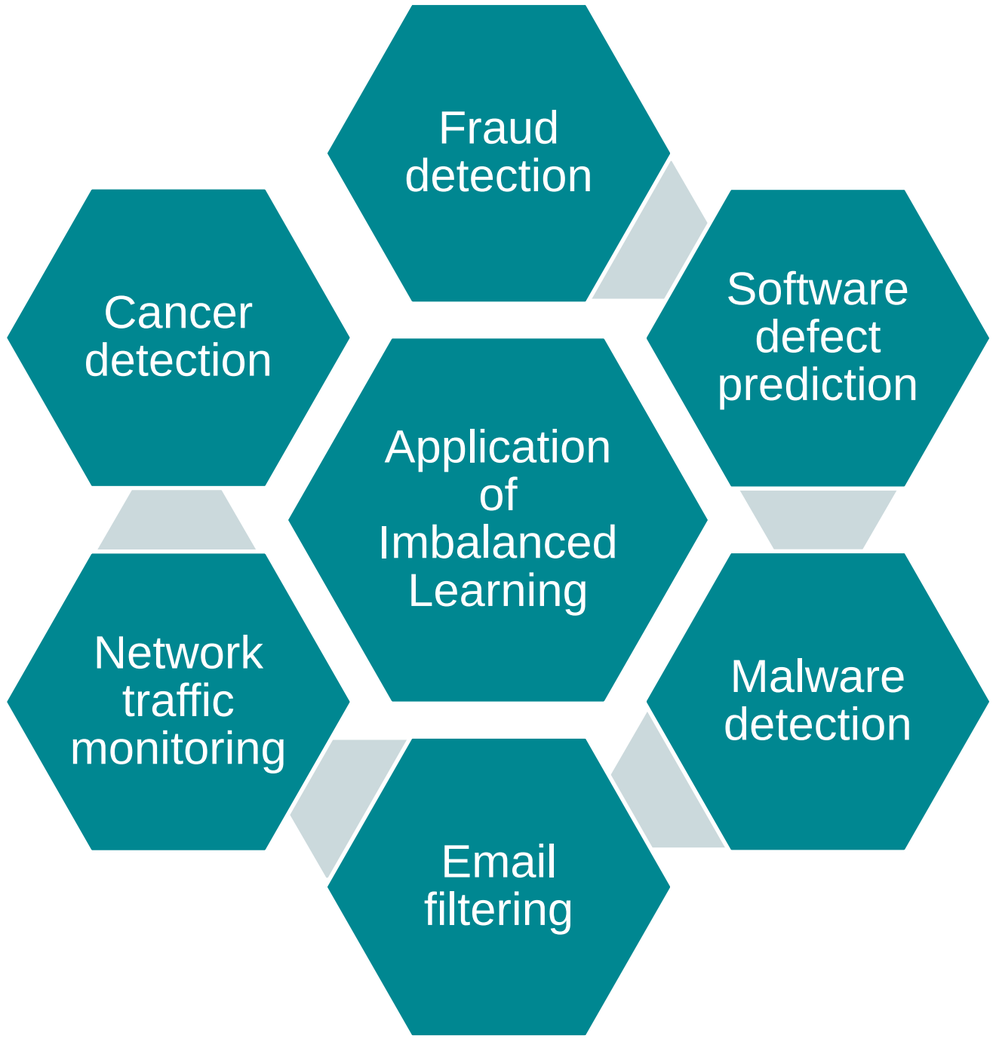
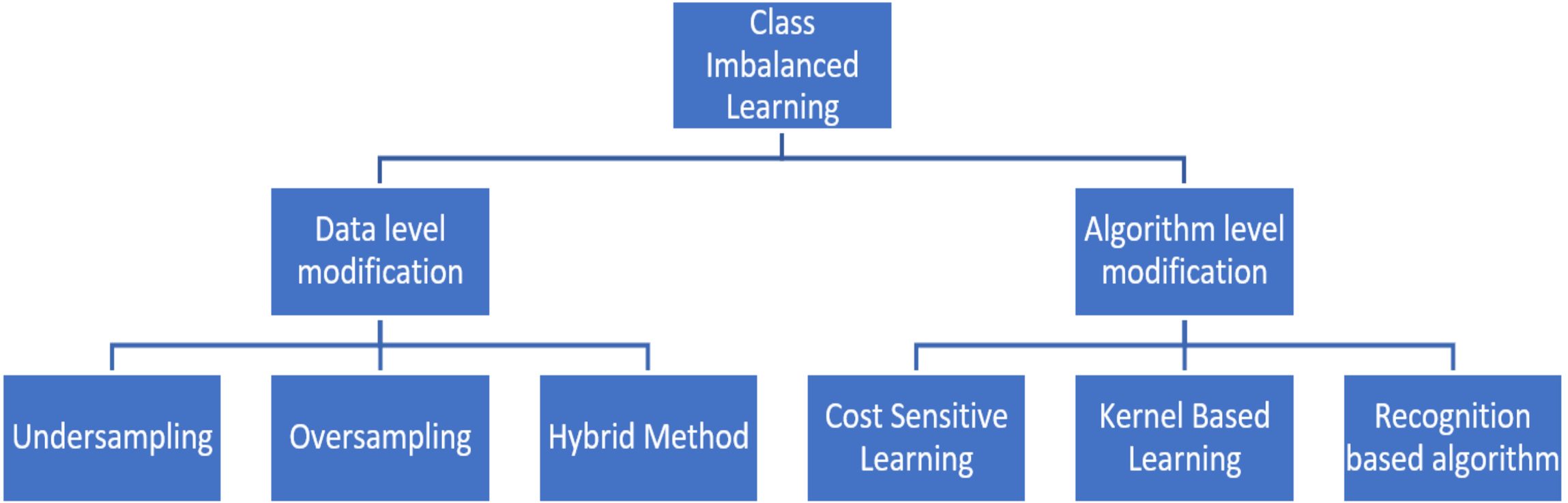


Figure: Sample Imbalanced datasets

Practical Application



Dealing with imbalanced data



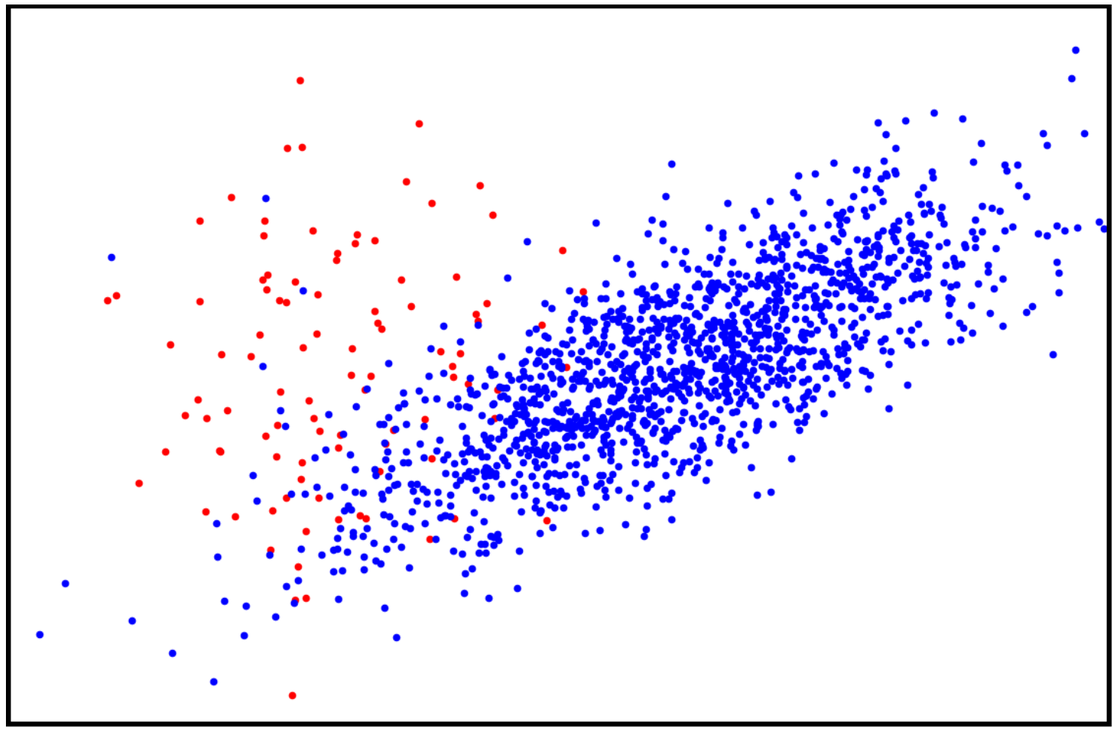
Data Level Modification

Less Complex

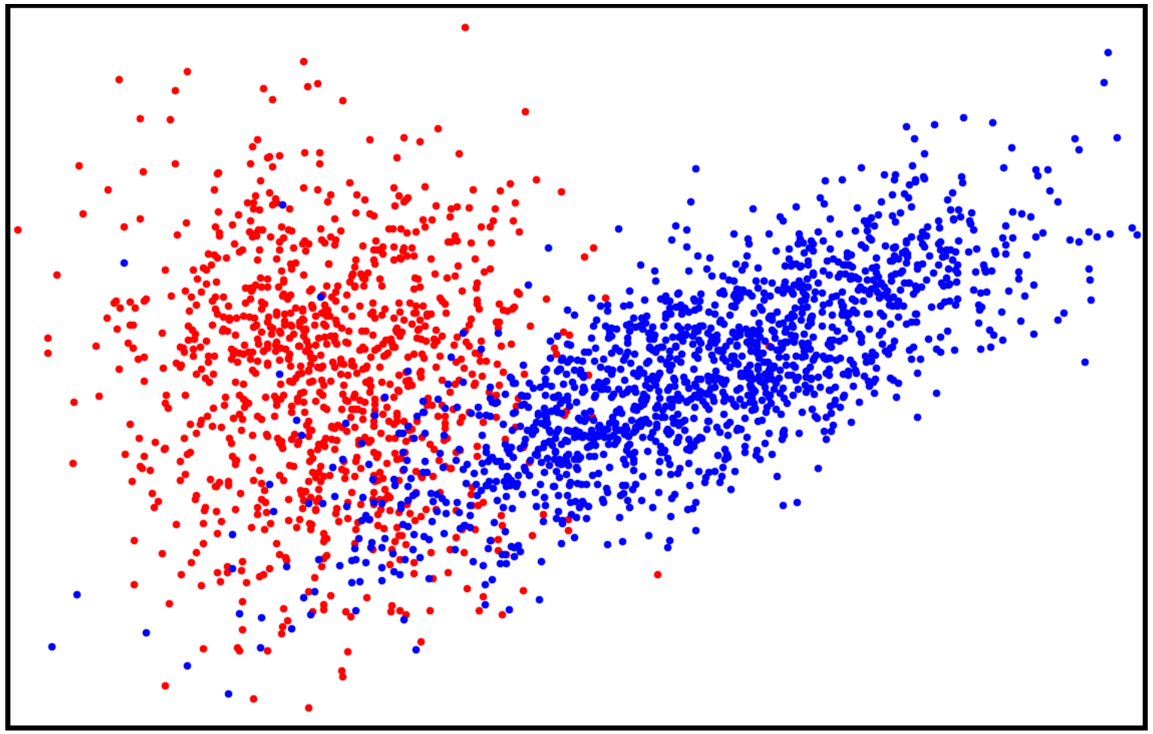
Loss of information

Not feasible in some cases

Oversampling

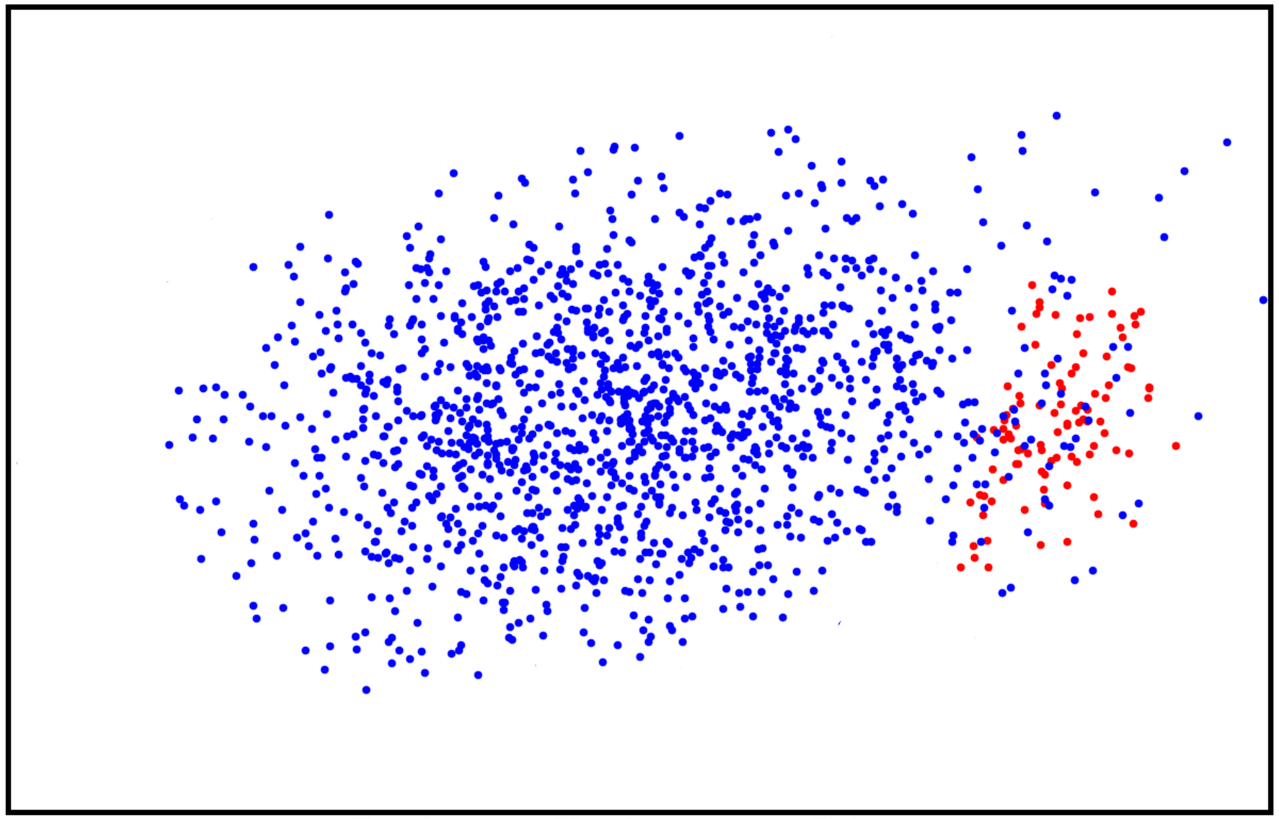


Before Oversampling

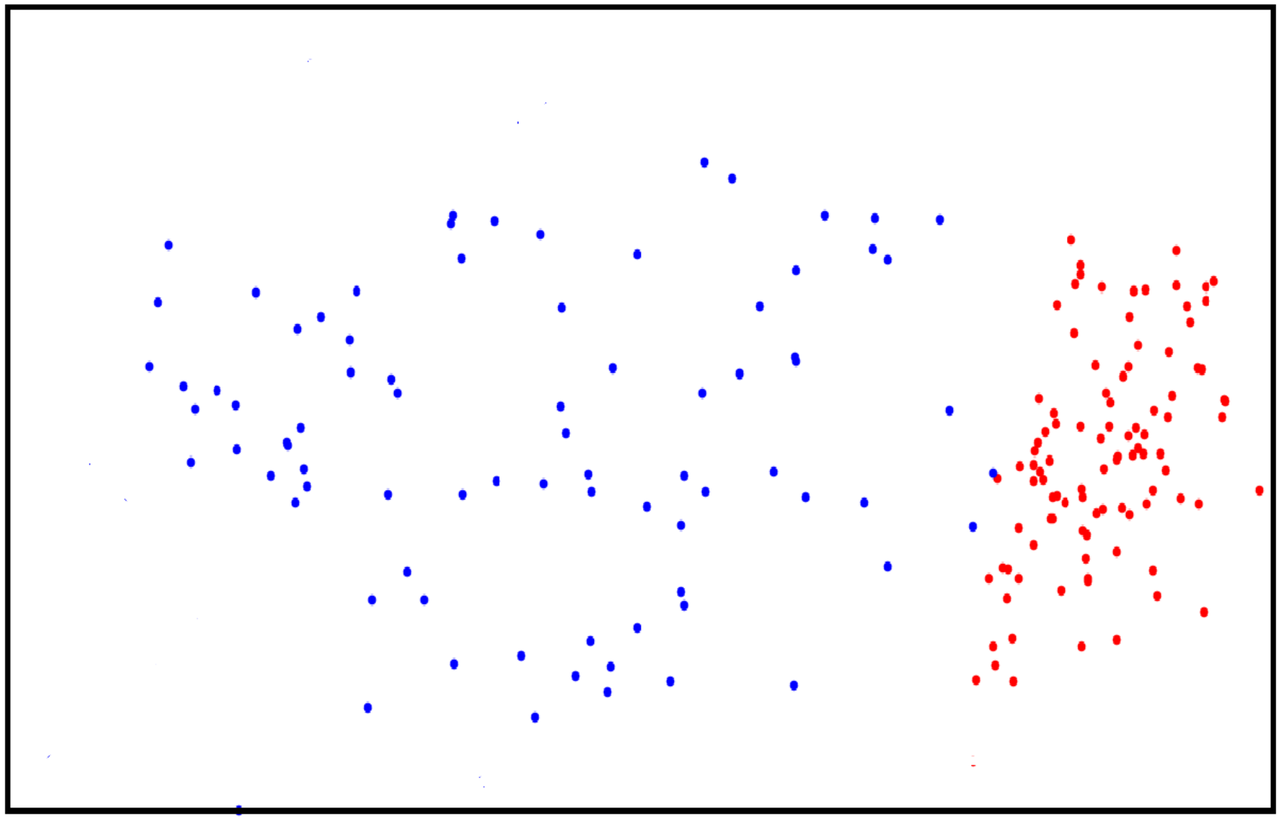


After Oversampling

Undersampling



Before Undersampling

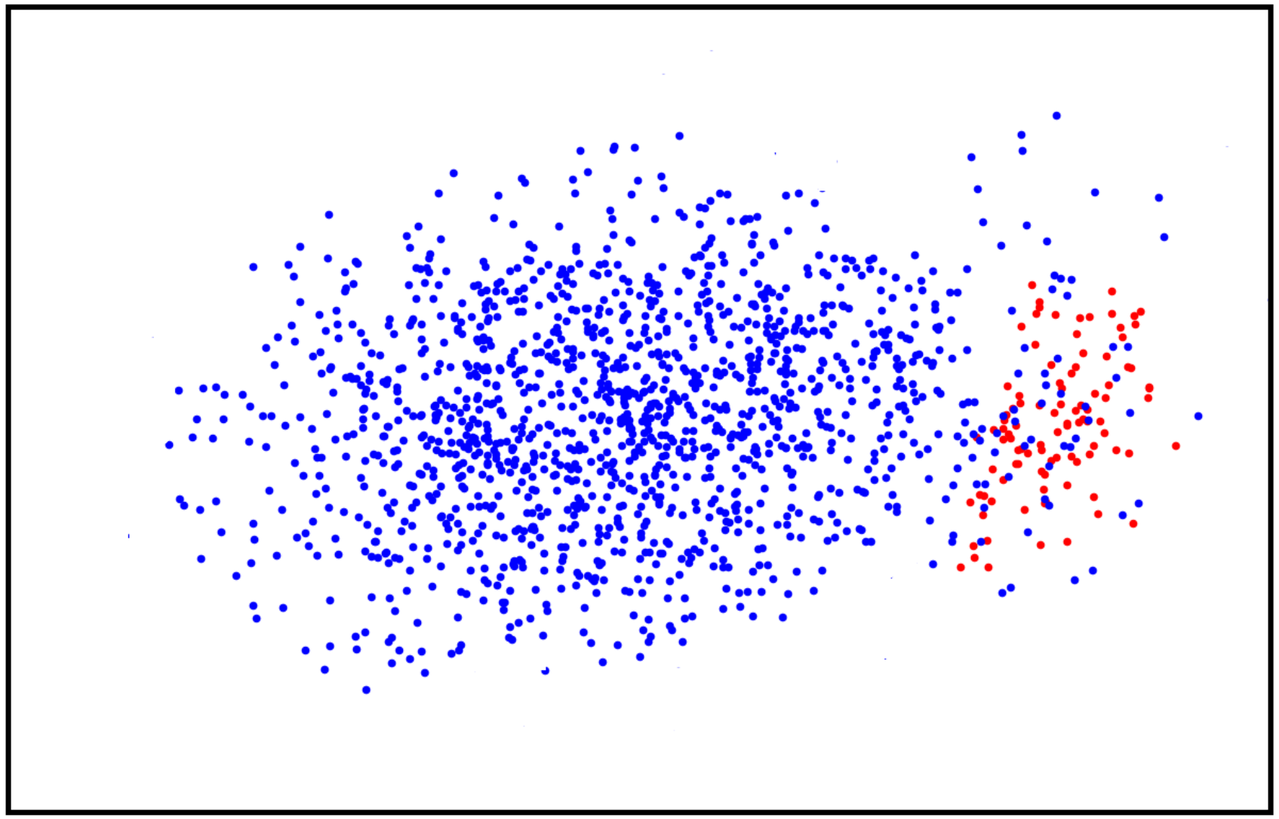


After Undersampling

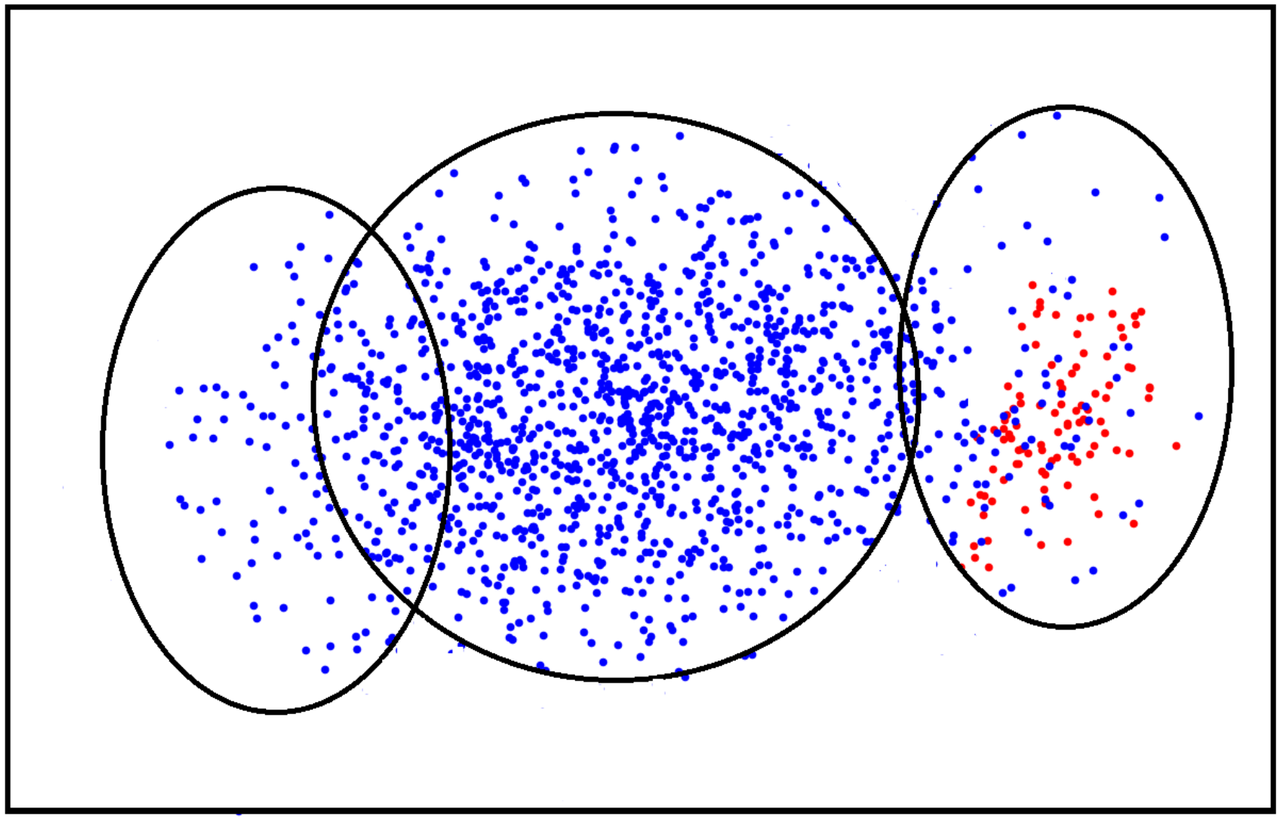
Proposed Approach

- Phase 1. In this phase we apply the incremental clustering algorithm to calculate clusters in the data set.
- Phase 2. Using outcomes of Phase 1, we apply undersampling inside some clusters and a supervised training model is created for each cluster.

Proposed Approach



Before Undersampling



Partial Undersampling

Clustering

- Modified global k-means algorithm (MGKM) is used for clustering.
- MGKM is not sensitive to the initial choice of cluster centre.
- MGKM calculates cluster centres at each incremental step.

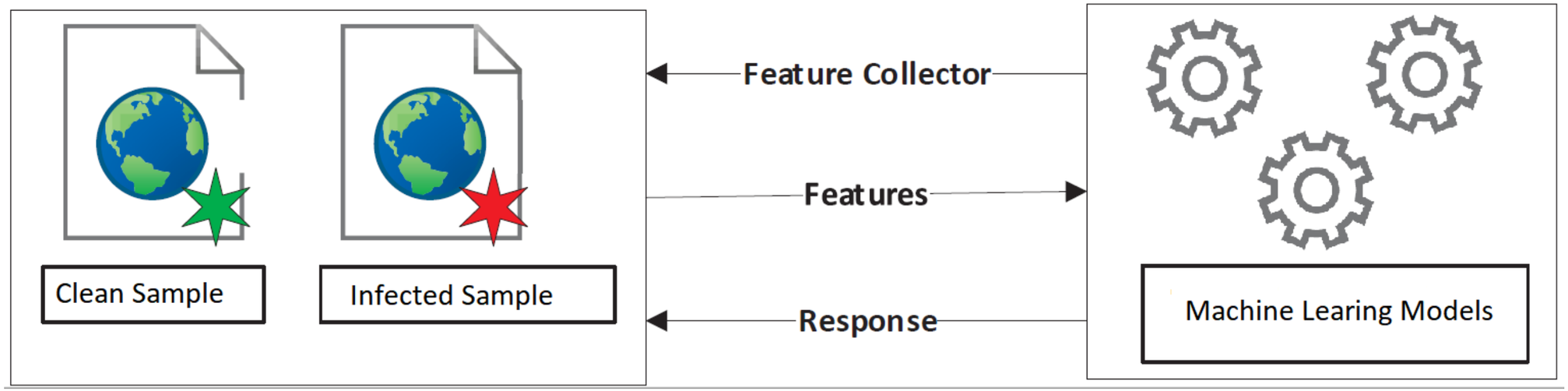
Types of cluster

- Type 1: Cluster containing points only from minority class.
- Type 2: Cluster containing points only from majority classes.
- Type 3: Cluster containing points from more than one class.

Classification Rules

1. For type 1 cluster, the observation is assigned to corresponding class.
2. For type 2 if the combined model is not trained, then the observation is assigned to majority class.
3. For type 2 if the combined model is trained, then the model is used to predict the observation.
4. For type 3, the trained model is used to predict the observation.

General Architecture



Experimental Setup

Dataset used:

- Us-Crime
- Ecoli
- Libras Move

Classifiers used:

- Random Forest
- KNN
- Adaboost
- SVM

Cost Matrix

	prediction $y = 1$	prediction $y = 0$
label $h(x) = 1$	$C_{1,1} = 0$	$C_{0,1} = P$
label $h(x) = 0$	$C_{1,0} = 1$	$C_{0,0} = 0$

* 0 = Clean Sample and 1 = Infected Sample

Numerical Results

Dataset	Class	Rnd. Forest	KNN	Adaboost	SVM
Us-Crime	Majority	98.92	98.92	97.29	99.19
	Minority	30.00	26.67	43.33	36.67
	Overall	93.73	93.48	93.23	94.49
Ecoli	Majority	98.36	96.72	96.72	96.72
	Minority	14.29	85.71	28.57	85.71
	Overall	89.71	95.59	89.71	95.59
Libras Move	Majority	100.00	100.00	100.00	100.00
	Minority	20.00	40.00	80.00	80.00
	Overall	94.44	95.83	98.61	98.61

Performance of mainstream classifiers

Dataset	Class	Rnd. Forest	KNN	Adaboost	SVM
Applying Random Undersampling (RUS)					
Us Crime	Majority	81.30	80.22	80.49	82.38
	Minority	86.67	93.33	83.33	93.33
	Overall	81.70	81.20	80.70	83.21
Ecoli	Majority	83.61	78.69	83.61	85.25
	Minority	57.14	100.00	71.43	85.71
	Overall	80.88	80.88	82.35	85.29
Libras Move	Majority	88.06	91.04	71.64	95.52
	Minority	100.00	100.00	80.00	100.00
	Overall	94.44	91.67	72.22	95.83
Applying SMOTE					
Us-Crime	Majority	94.04	82.11	91.06	90.24
	Minority	63.33	80.00	60.00	83.33
	Overall	91.73	81.95	88.72	89.72
Ecoli	Majority	95.08	91.80	93.44	90.16
	Minority	85.71	85.71	71.43	85.71
	Overall	94.12	91.18	91.18	89.71
Libras Move	Majority	100.00	95.52	100.00	97.01
	Minority	60.00	100.00	80.00	80.00
	Overall	97.22	95.83	98.61	95.83
Applying our proposed method					
Us-Crime	Majority	89.70	87.26	83.74	87.80
	Minority	73.33	76.67	83.33	90.00
	Overall	88.47	86.47	83.71	87.97
Ecoli	Majority	95.08	91.80	86.89	93.44
	Minority	71.43	85.71	57.14	85.71
	Overall	92.65	91.18	83.82	92.65
Libras Move	Majority	94.03	86.57	85.07	92.54
	Minority	80.00	100.00	100.00	100.00
	Overall	93.06	87.50	98.61	93.06

Comparison of different methods

Future Works

- Introducing additional types of clusters
- Dynamically selecting number of clusters
- Applying more classification rules
- Designing sophisticated classifiers for imbalanced data

Thank you