Comparative approach and Ensemble learning on Multiple datasets

Imbalanced dataset

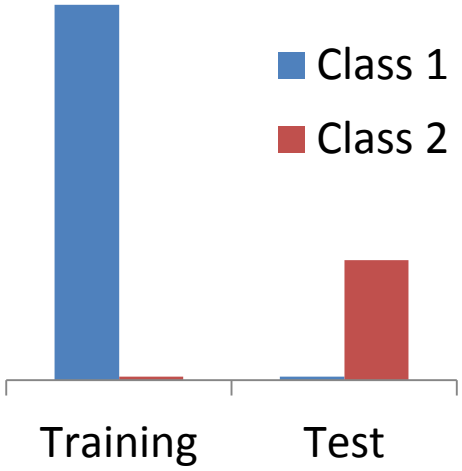
- Dataset in which amount of training and test samples are skewed towards one class

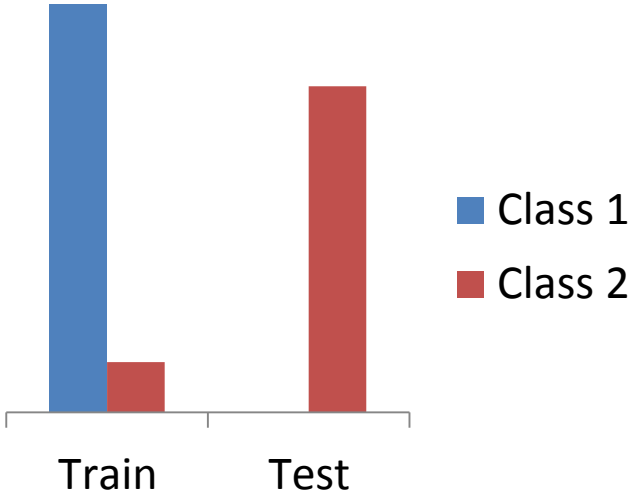
Wine data*	Class1	Class2
Training set	4850	50
Test set	48	1549

*UCI Machine Learning Repository

CCS Chall-enge data+	Class1	Class2
Training set	9382	618
Test set	0	4000

+mlcomp repository





Classification using single classifier

- Data preprocessing: Scaling train data and test data

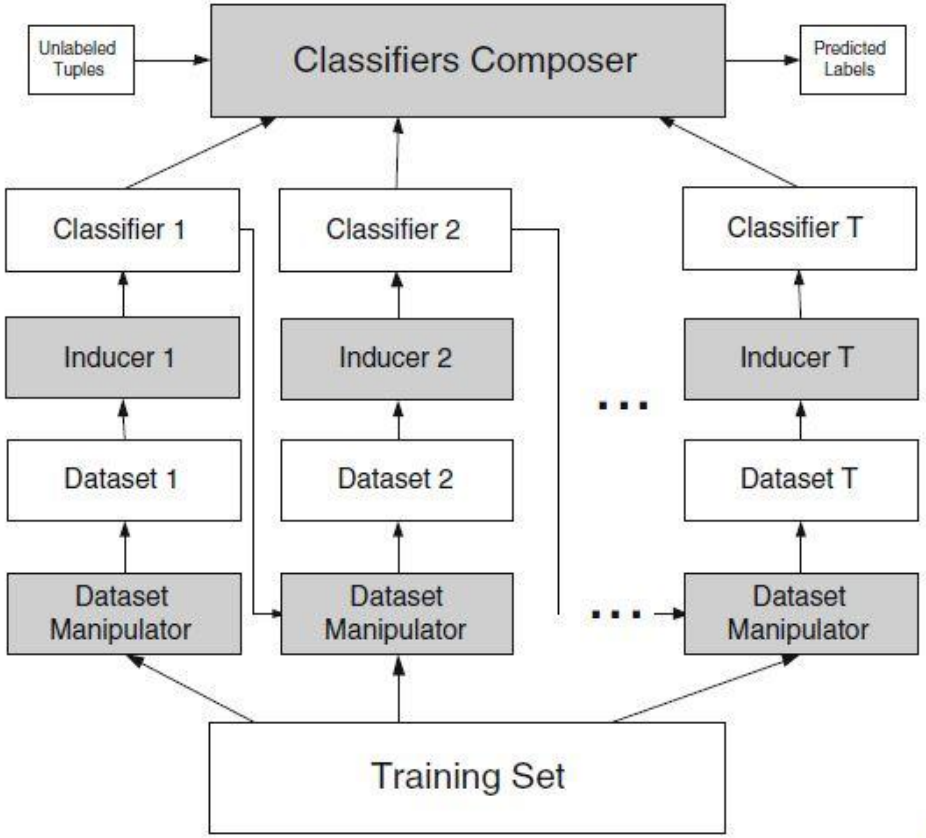
  Train: [-10,10] → [-1,+1]

  Test: [-11,+8] → [-1.1,0.8]

	Accuracy on Wine data			
Classifier (Inducer)	Training (without scaling)	Training (with scaling)	Test (without scaling)	Test (with scaling)
SVM (SMO)	99.55 %	99.44%	24.17%	29.43%
Logistic Regression	99.69%	99.65%	61.05%	21.54%
Neural Network (BPNN)	93.0%	30%	61%	60%

What is the need for Ensemble learning

- Adaboost: Cascade combination of weak classifiers
- Up sampling of samples from minor class and Down sampling of samples from major class



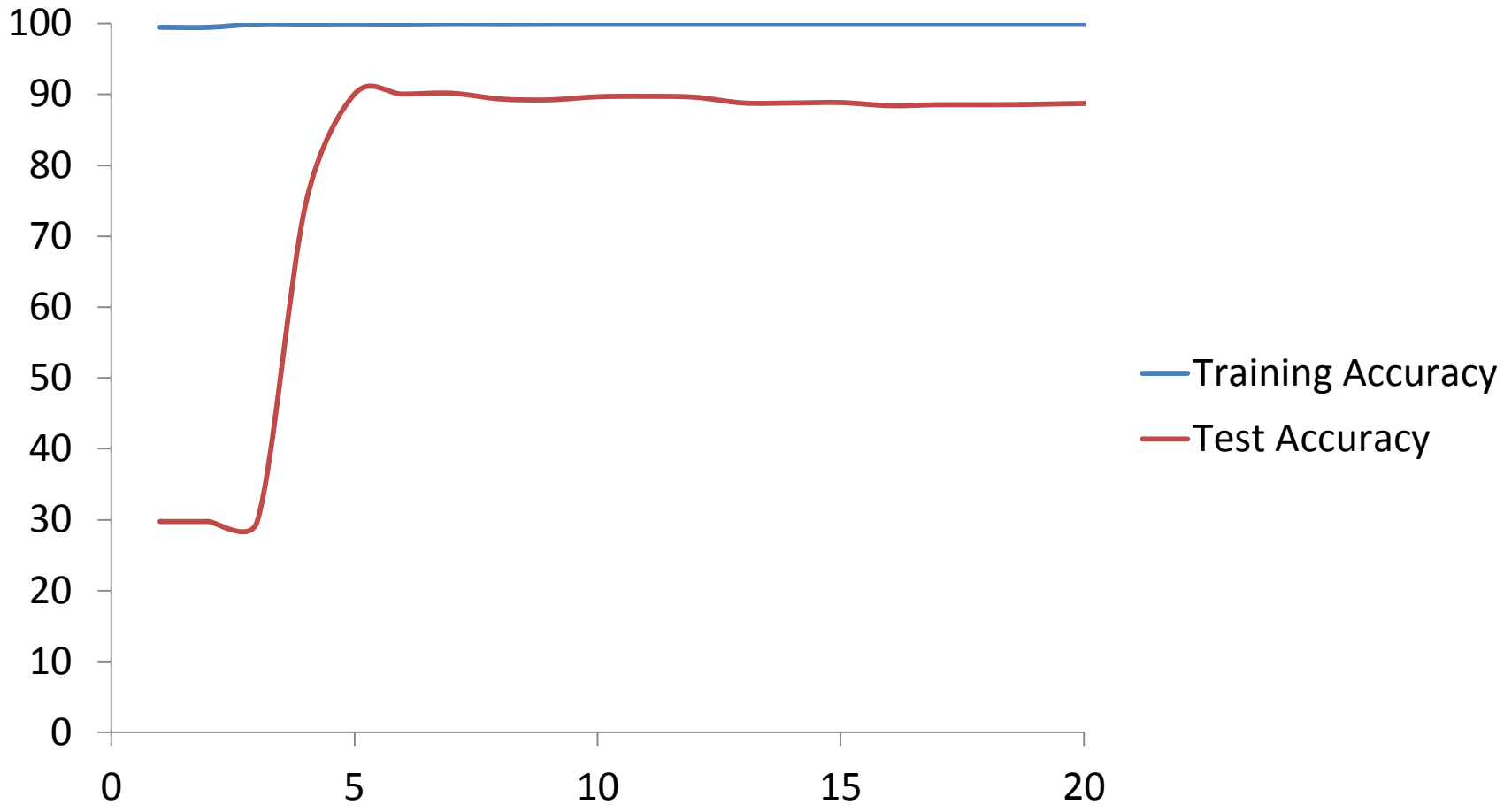
Adaboost Algorithm

Require: I (a weak inducer), T (the number of iterations), S (training set)
Ensure: $M_t, \alpha_t; t = 1, \ldots, T$

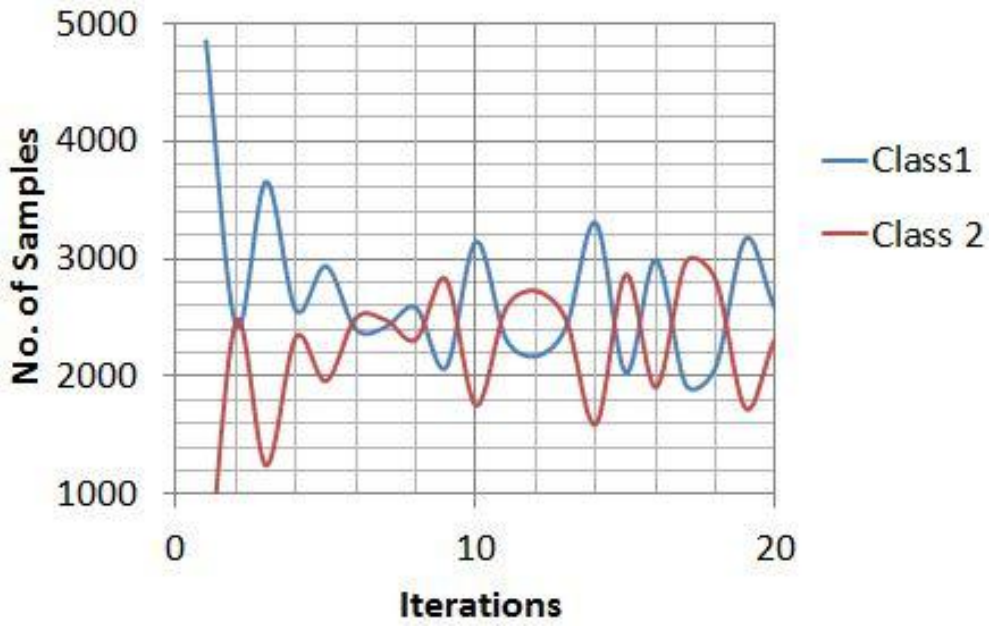
1: $t \leftarrow 1$
2: $D_1(i) \leftarrow 1/m; i = 1, \ldots, m$
3: **repeat**
4: Build Classifier M_t using I and distribution D_t
5: $\varepsilon_t \leftarrow \sum_{i:M_t(x_i) \neq y_i} D_t(i)$
6: **if** $\varepsilon_t > 0.5$ **then**
7: $T \leftarrow t - 1$
8: exit Loop.
9: **end if**
10: $\alpha_t \leftarrow \frac{1}{2} \ln(\frac{1-\varepsilon_t}{\varepsilon_t})$
11: $D_{t+1}(i) = D_t(i) \cdot e^{-\alpha_t y_t M_t(x_i)}$
12: Normalize D_{t+1} to be a proper distribution.
13: $t++$
14: **until** $t > T$

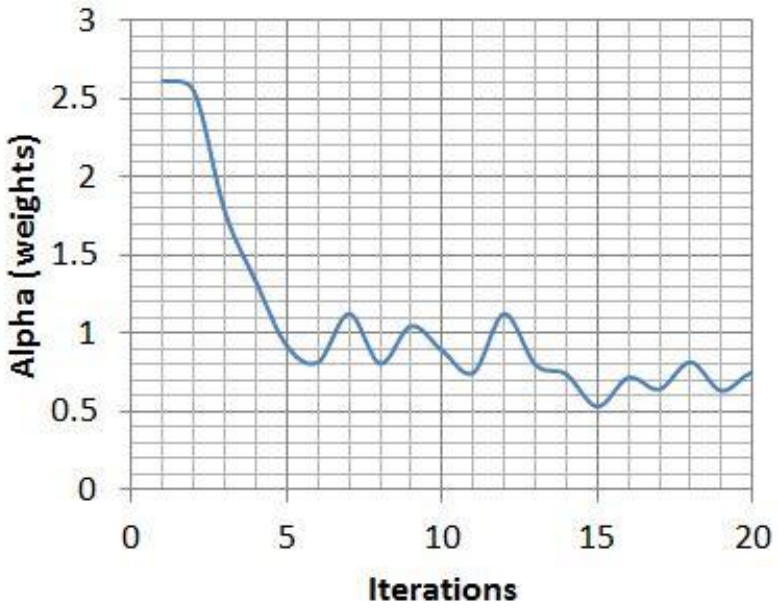
$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot M_t(x)\right)$$

Performance at each iteration



Interpretation of adaboost





Final results

Classifier (Inducer)	Accuracy of single inducer (wine data)	Accuracy of boosted classifier (wine data)
SVM (SMO)	29.43%	89%
Logistic Regression	61.05%	87%

Classifier (Inducer)	Accuracy of single inducer (ccs challenge data)	Accuracy of boosted classifier (ccs challenge data)
SVM (SMO)	2%	6%