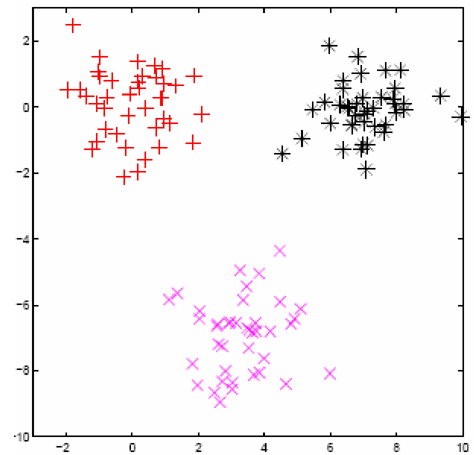
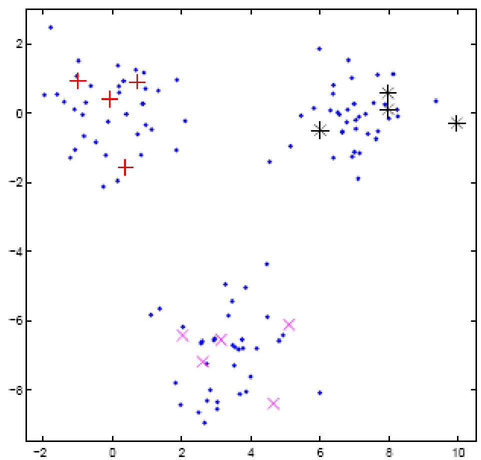
Semi-supervised Learning

Anil K. Jain
(with Rong Jin, Pavan Mallapragada and Yi Liu)
Department of Computer Science and Engineering
Michigan State University

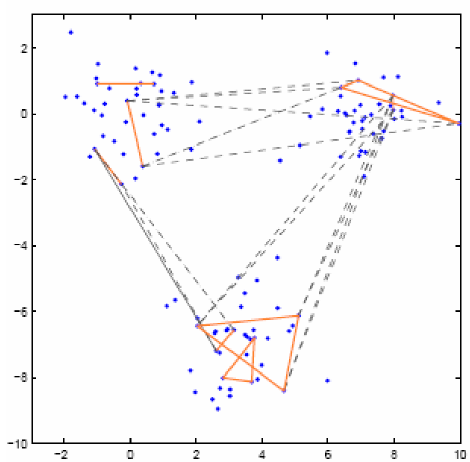
Semi-supervised Learning



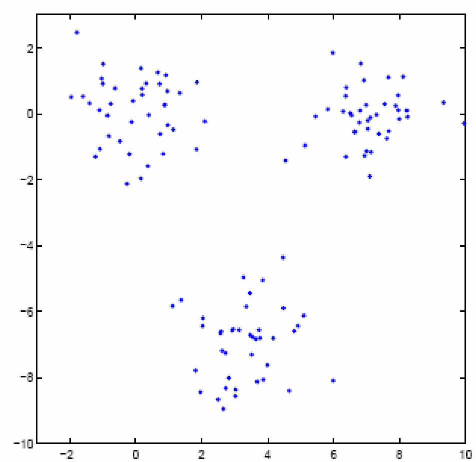
Supervised



Partially Labeled



Pairwise Constraints



Unsupervised

Why Semi-supervised Learning?

- Data labeling is expensive and difficult
  - Remote sensing
  - Labeling large images at pixel level
- Labeling is often unreliable
  - Disagreement among experts
- Unlabeled examples
  - Easy to obtain in large numbers
  - e.g. webpage classification, bioinformatics, nondestructive inspection, image classification

Problem

- Classification
  - Use unlabeled data to improve classifier performance (SemiBoost)
- Clustering
  - Use labeled points or pairwise constraints to find natural groupings (BoostCluster)

No. of labeled points << no. of unlabeled points

Chapelle, Scholkopf and Zien (eds.), Semi-Supervised Learning, 2006

Is unlabeled data useful?

- In general yes, but not always
- Classification error reduces
  - Exponentially with labeled examples
  - Linearly with unlabeled examples

  (Castelli and Cover, IEEE Inf. Th., 1996)
- Capacity of labeled samples
  - How many unlabeled points can a given labeled set accommodate?
- Several specialized semi-supervised learning algorithms are available

SemiBoost

- Improve the performance of any supervised classifier using unlabeled data
- Graph based approach defines consistency between similarity matrix and assigned labels
- Boosting allows us to incorporate the given classifier in minimizing the objective function

Boosting

- Improve the performance of a supervised classifier
- Train successive component classifiers with a subset of unlabeled samples that is "most informative"; use the ensemble classifier
- AdaBoost
  - Use true labels to select the subset
- SemiBoost
  - Define "consistency" of unlabeled samples to select the subset and to assign class labels

Objective Function

Unlabeled samples close to each other have similar labels; unlabeled samples near labeled samples share the same label; S= similarity matrix

- Unlabeled sample energy

$$F_u(\mathbf{y}_u, S) = \sum_{ij} S_{i,j} \exp(y_i^u - y_j^u)$$

"Exponential linear" in y^u

- Labeled sample energy

$$F_l(\mathbf{y}, S) = \sum_{i=1}^{n_l} \sum_{j=1}^{n_u} S_{i,j} \exp(-2 y_i^l y_j^u).$$

- Minimize total energy

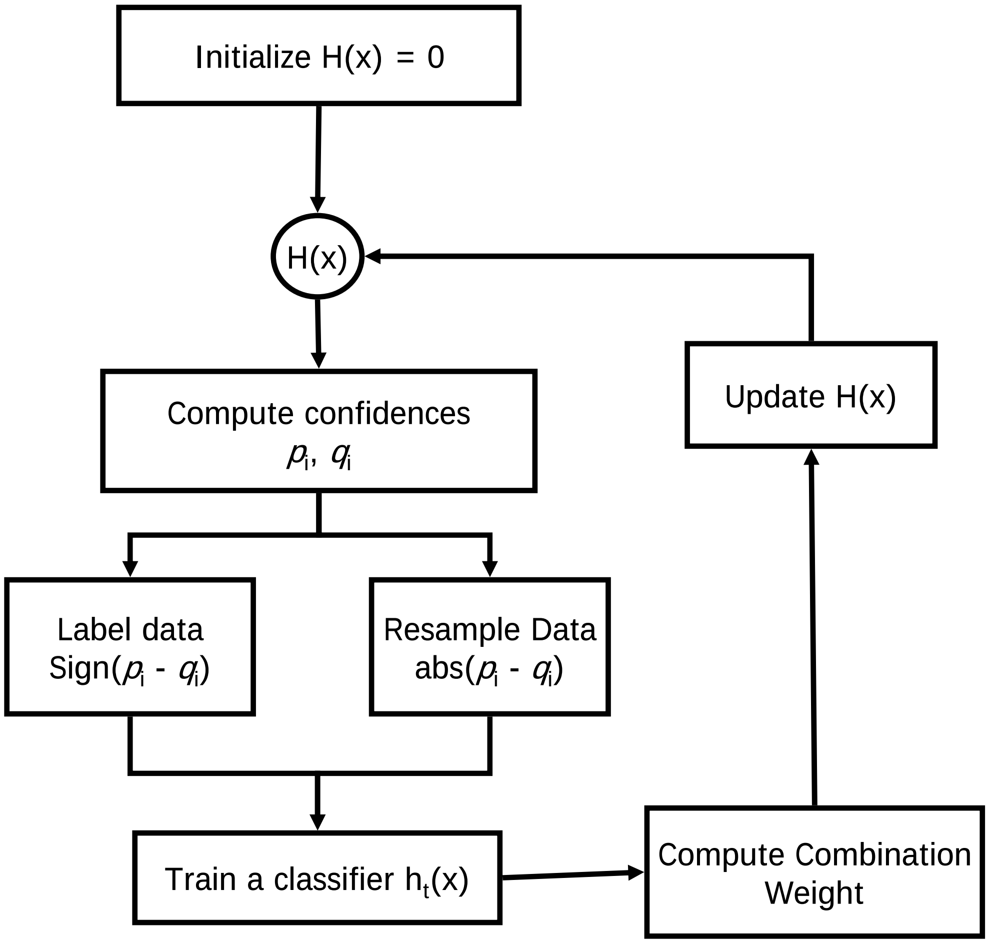
$$F(\mathbf{y}, S) = F_l(\mathbf{y}, S) + C F_u(\mathbf{y_u}, S).$$

C is the ratio of no. of labeled samples to no. of unlabeled samples

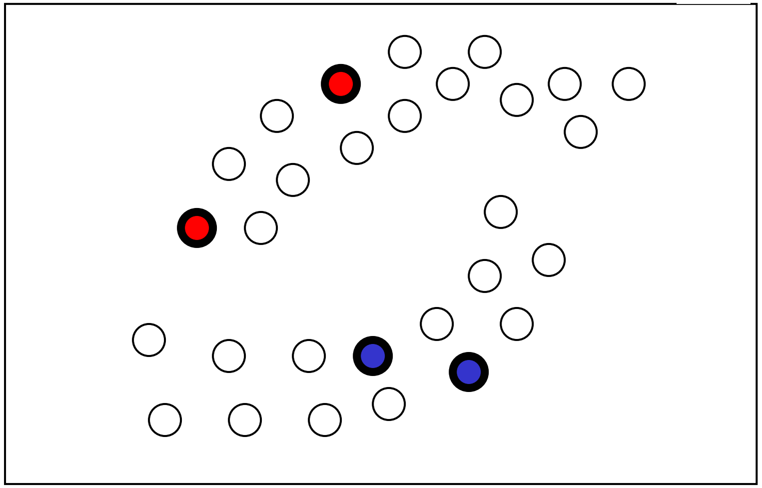
Solution

- Replace y^u in the energy function with an ensemble classifier prediction
- Form of component classifier is given (decision tree, SVM)
- Use boosting to learn component classifiers and weights
- Output is a classifier that learns from both labeled and unlabeled examples

SemiBoost Algorithm

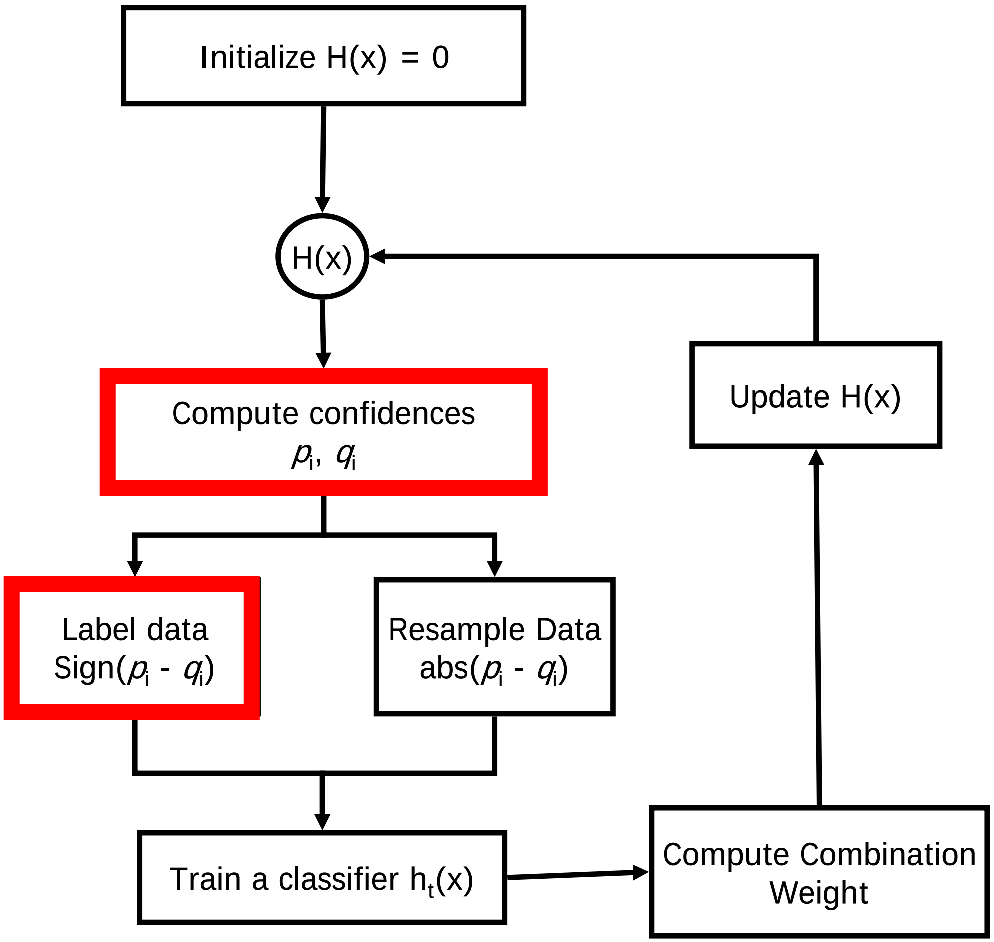


Iteration 1

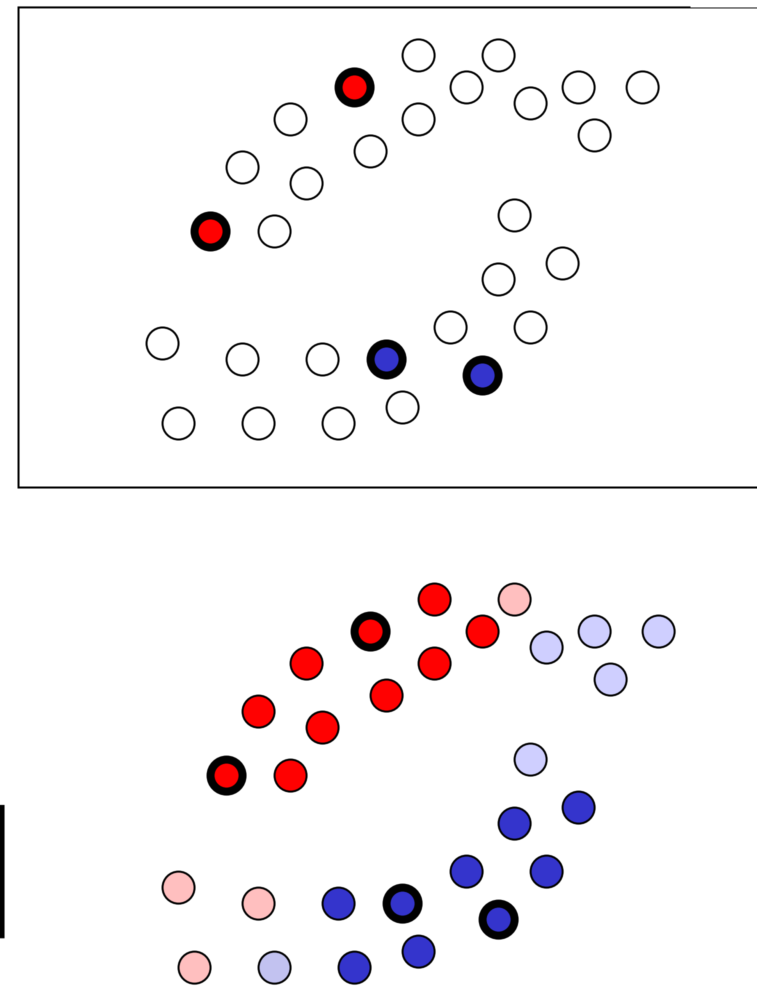


H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm



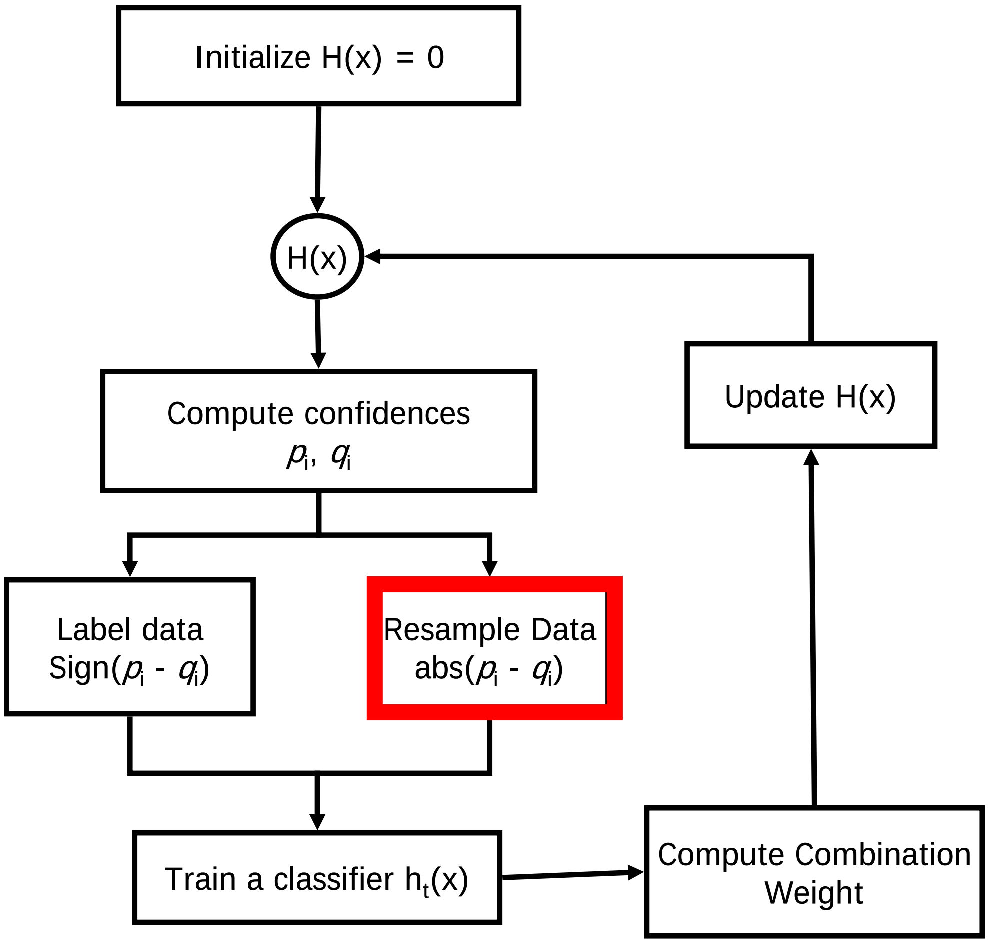
Iteration 1

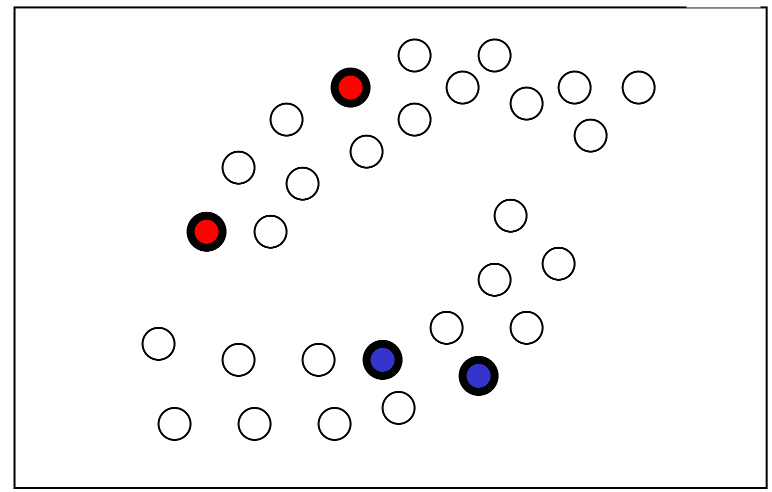


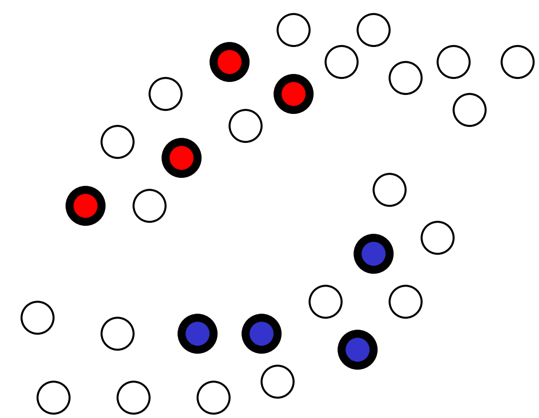
H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm

Iteration 1

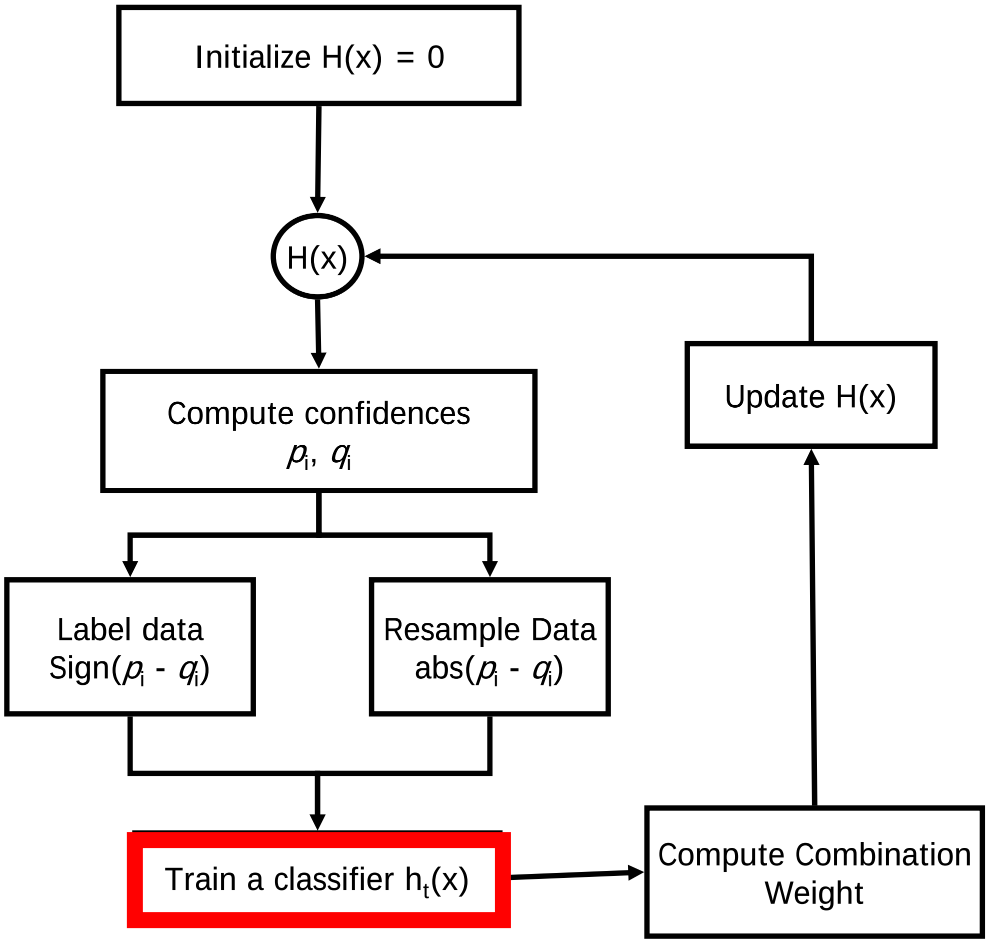




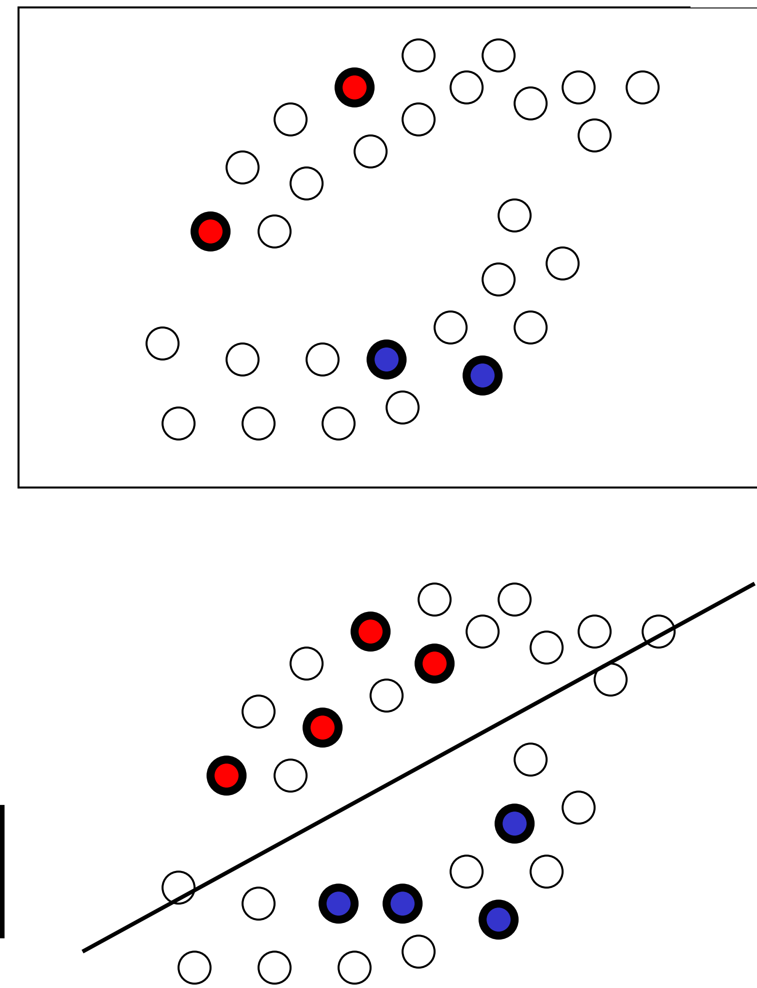


H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm

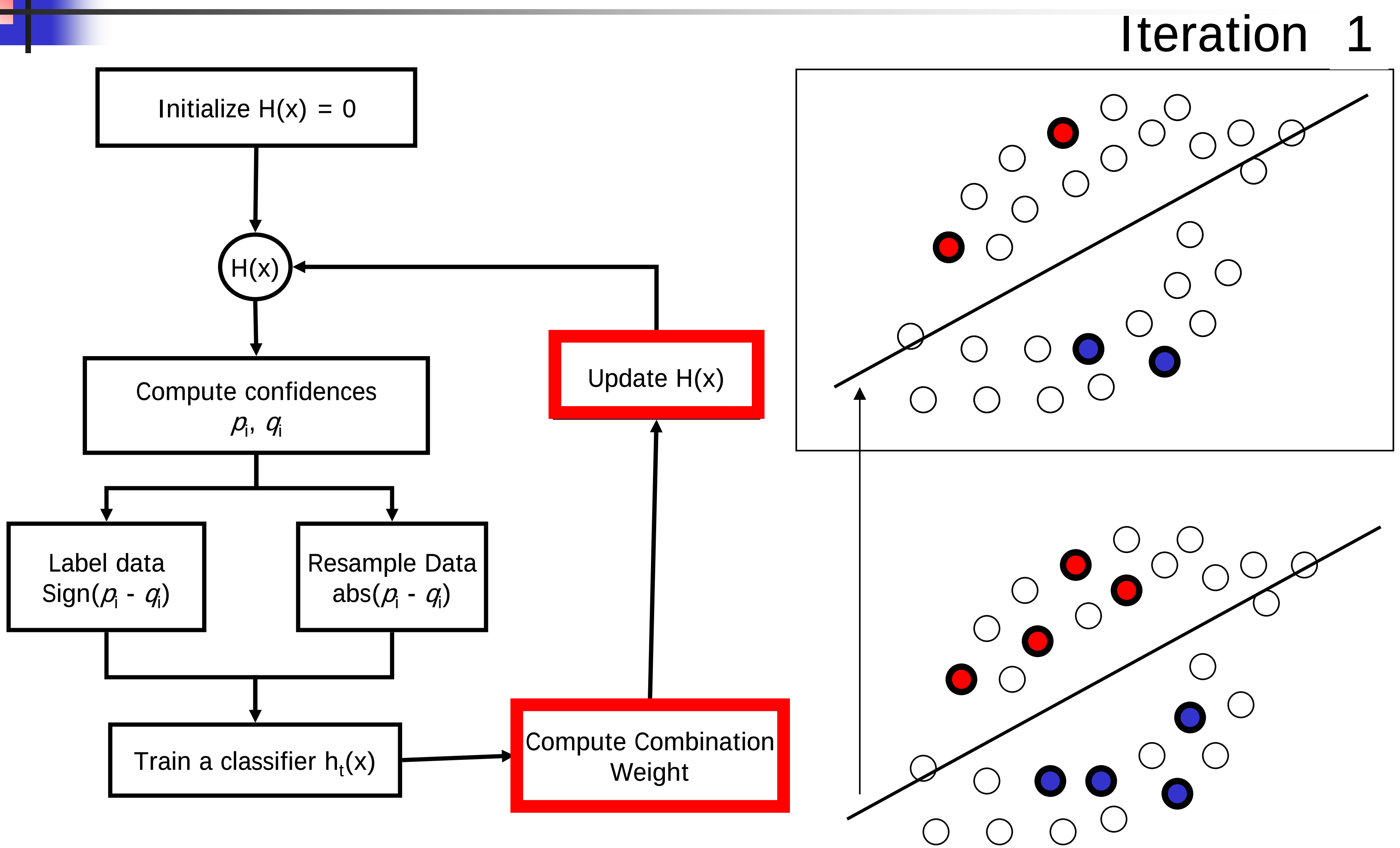


Iteration 1



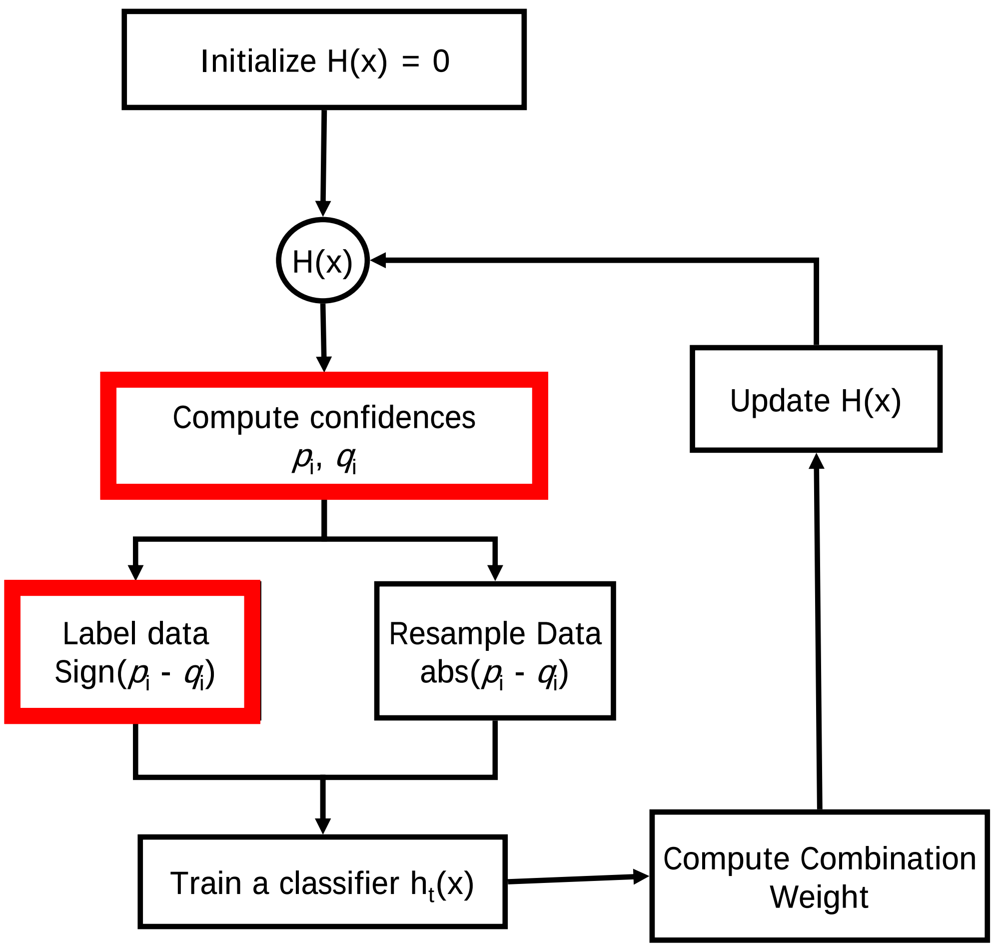
H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm

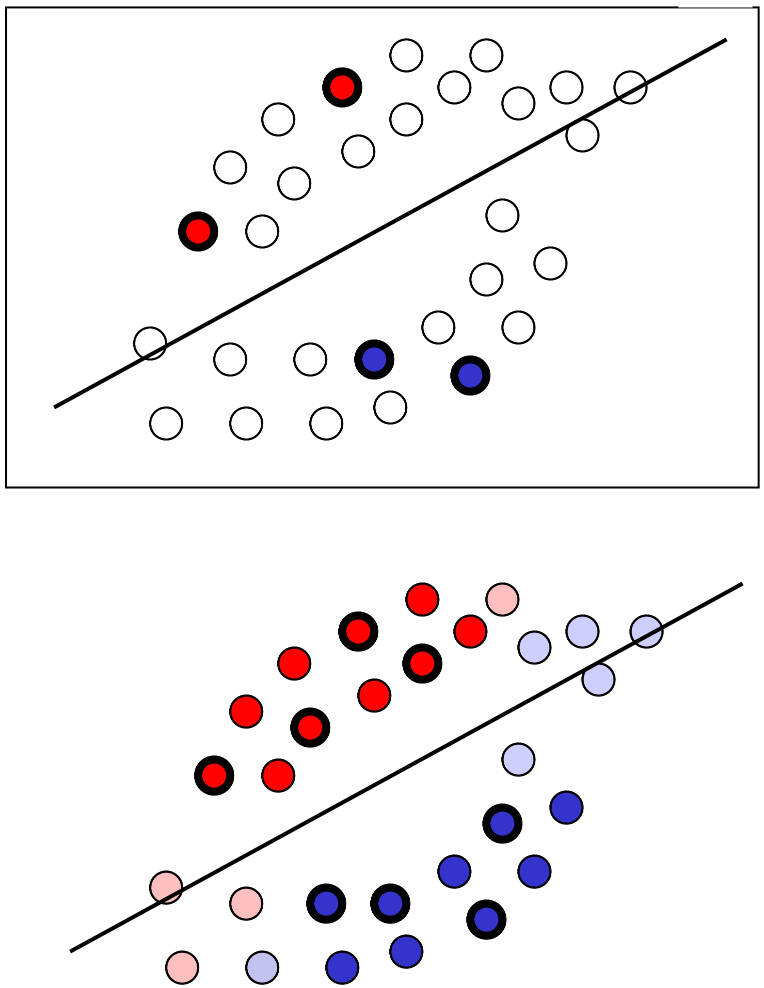


H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm

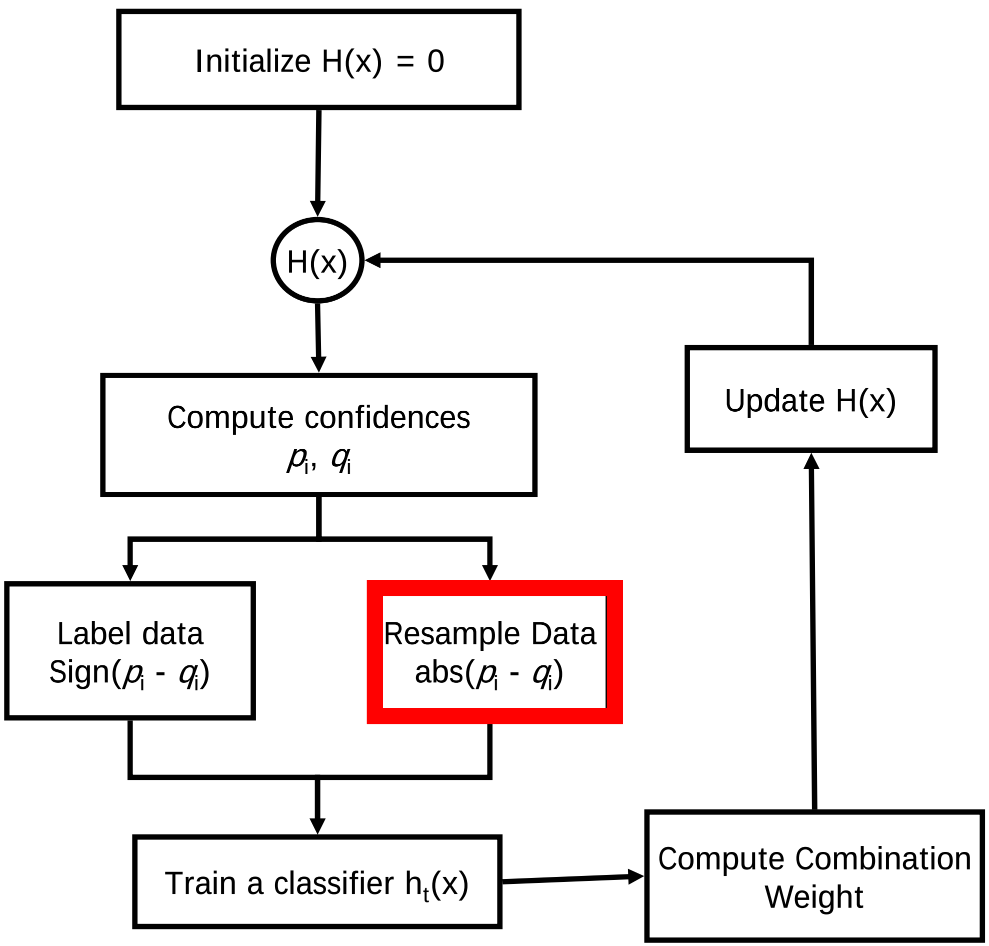


Iteration 2

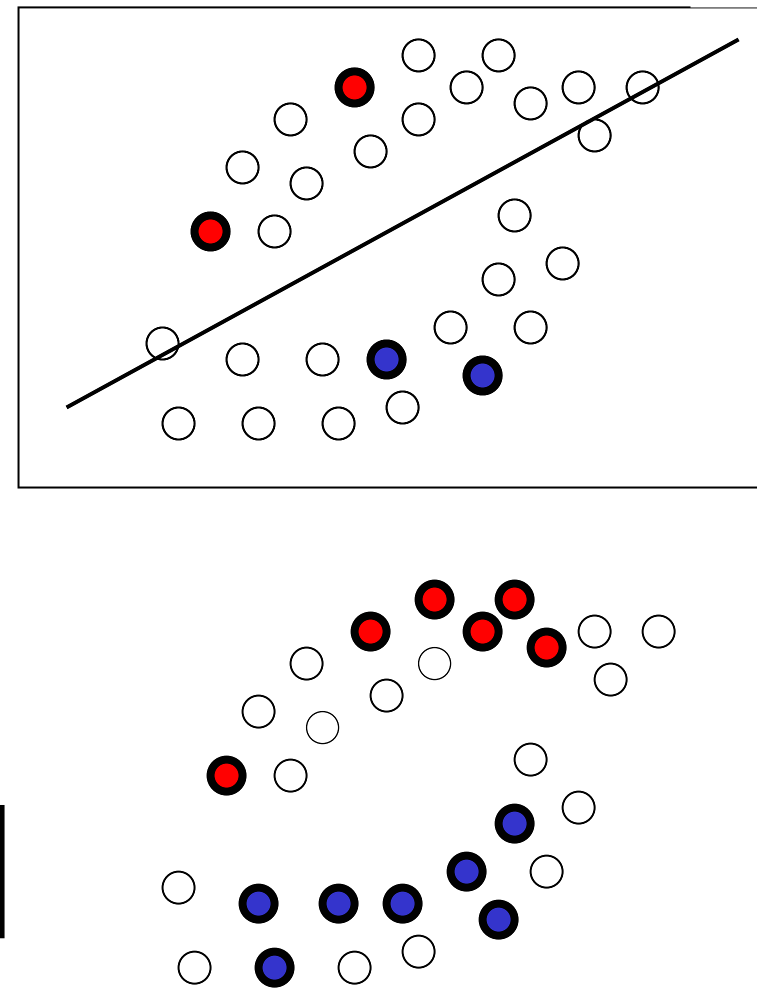


H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm



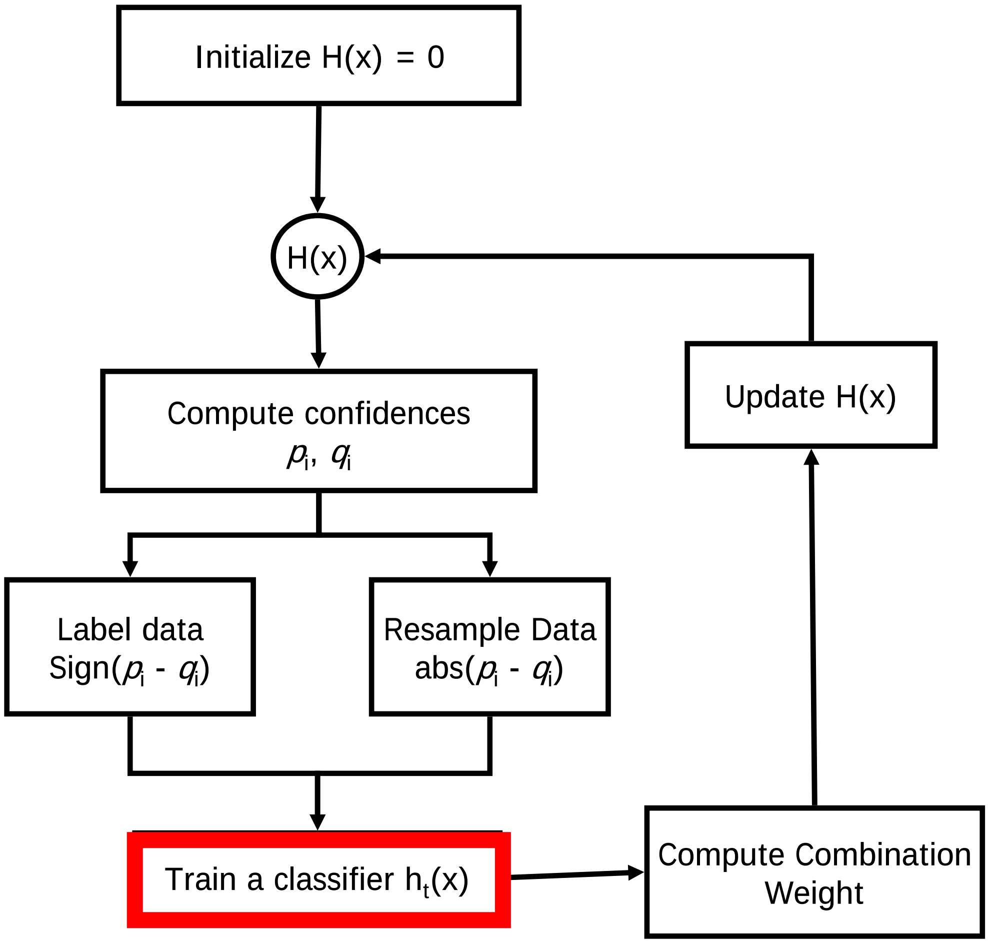
Iteration 2

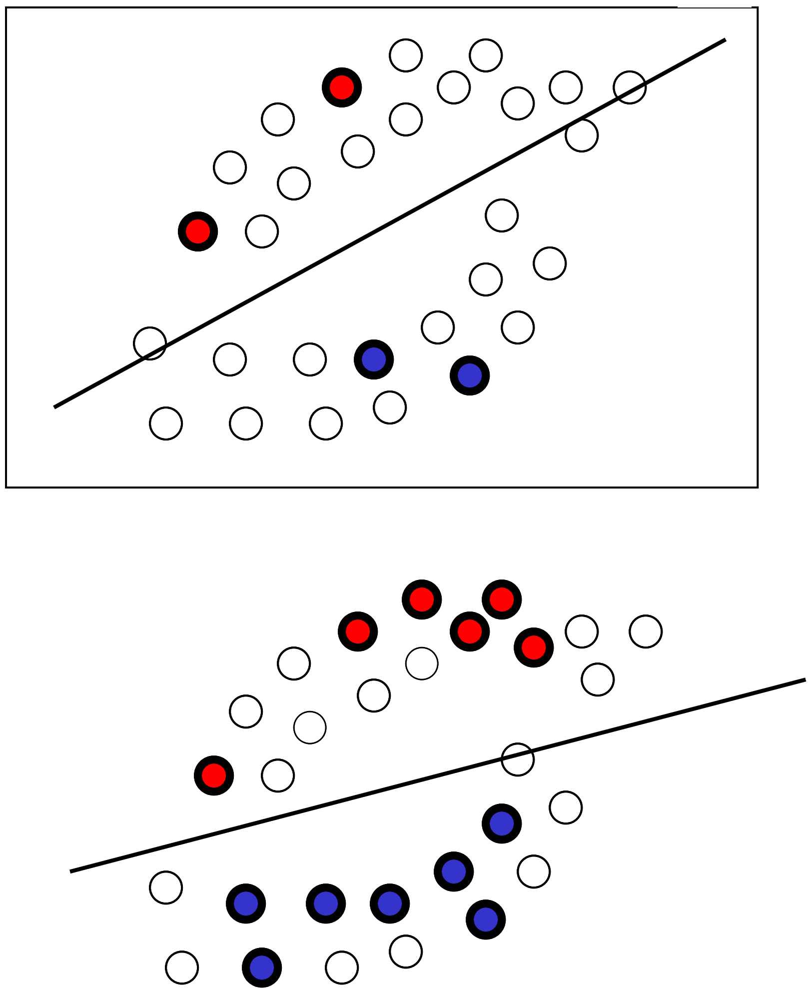


H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm

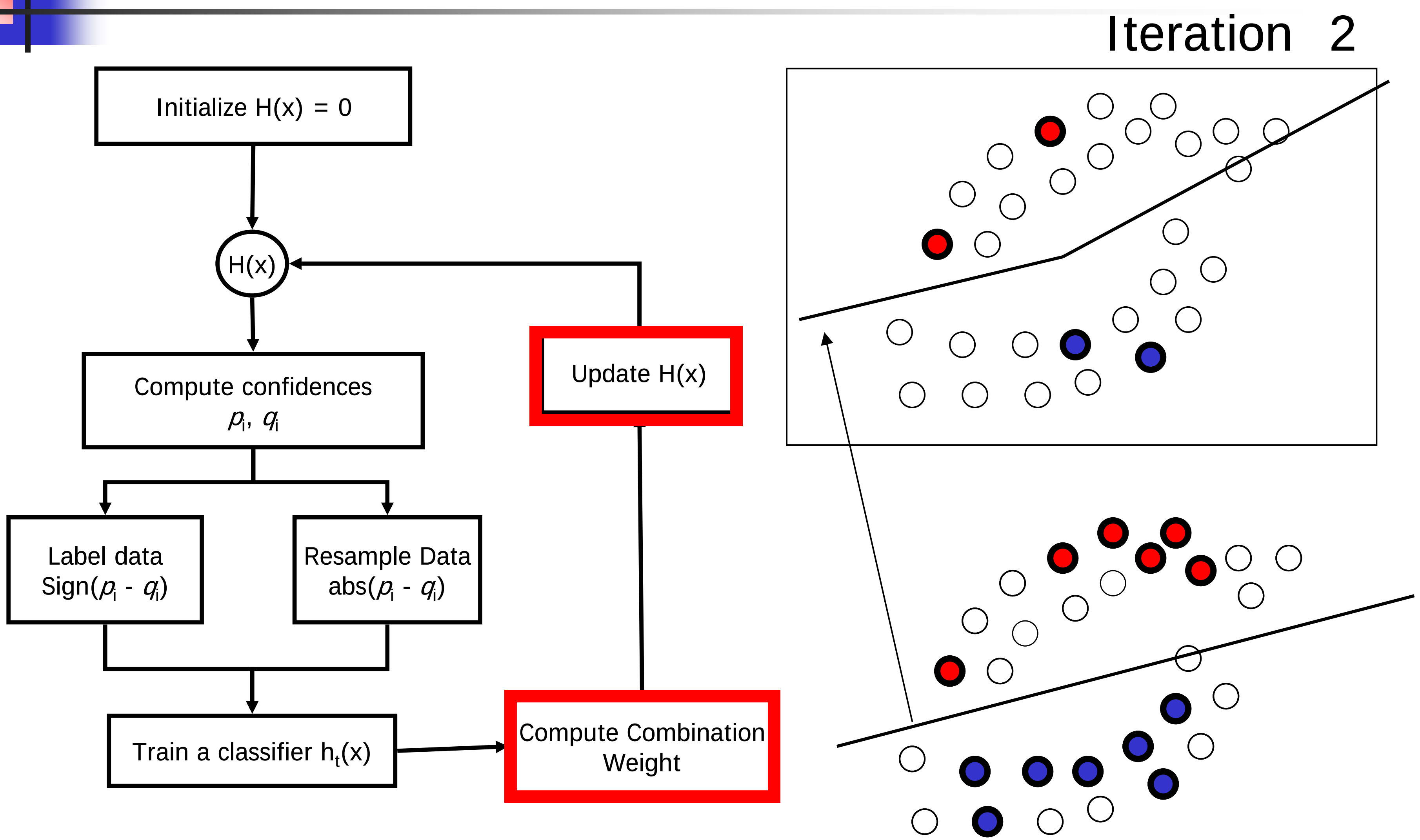
Iteration 2





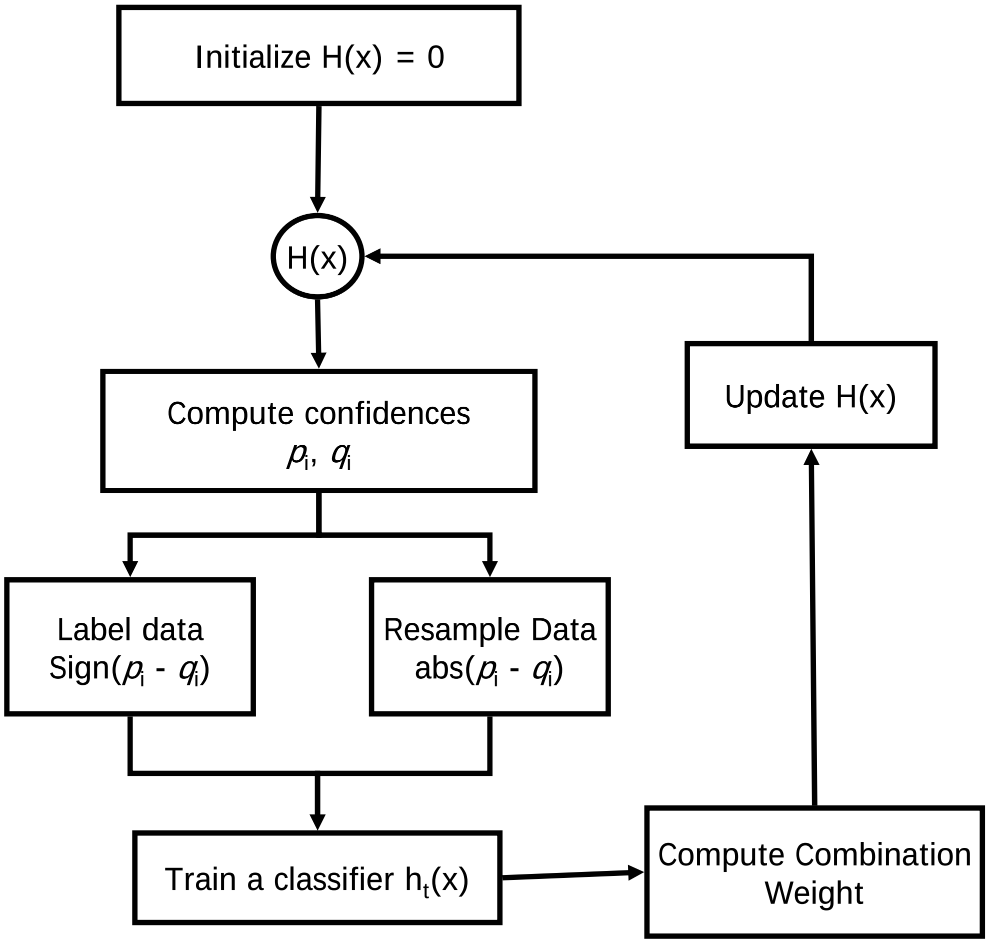
H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm

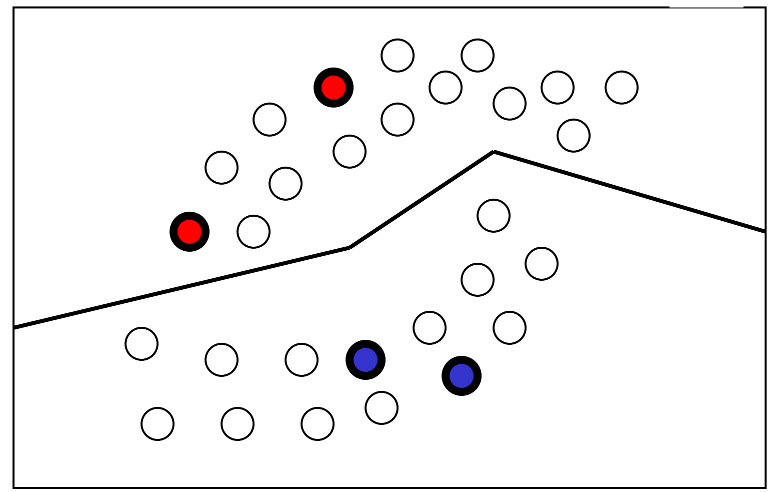


H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Algorithm



Iteration 3



H(x) is ensemble and h(x) is base classifier; p,q = confidence in assigning x to class 1 and 2

SemiBoost Performance

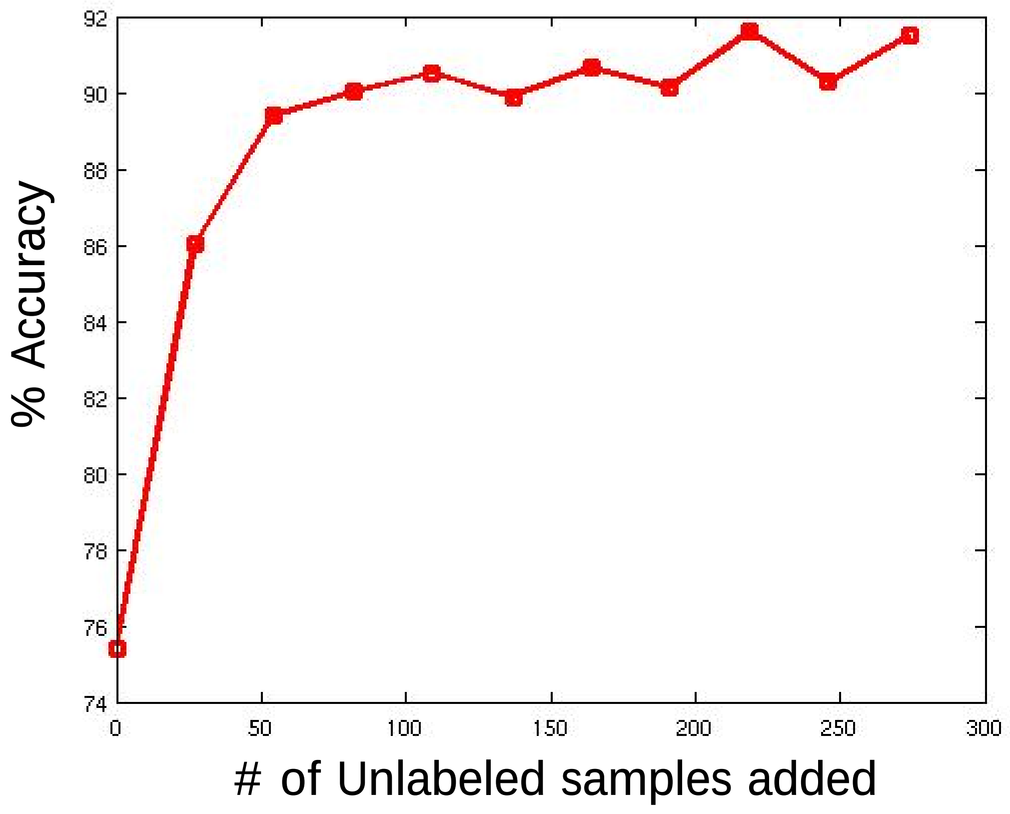
Dataset	n	d	SVM	SB-SVM
Wdbc	569	14	75.5 (5.7)	91.0 (3.5)
Isolet	600	51	90.8 (3.7)	94.8 (3.3)
Optdigits	1143	42	87.8 (2.3)	95.9 (2.6)
Heart	270	9	68.4(6.7)	77.7 (3.5)
Same-300	199	20	68.3 (6.5)	70.4 (9.1)

SVM is trained on 5 labeled samples per class; two most populated classes only; standard deviation based on runs with 10 different training sets of 5 samples/class

SemiBoost: Inductive Performance

wdbc (UCI dataset): 569 samples; 14 features; 2 classes;

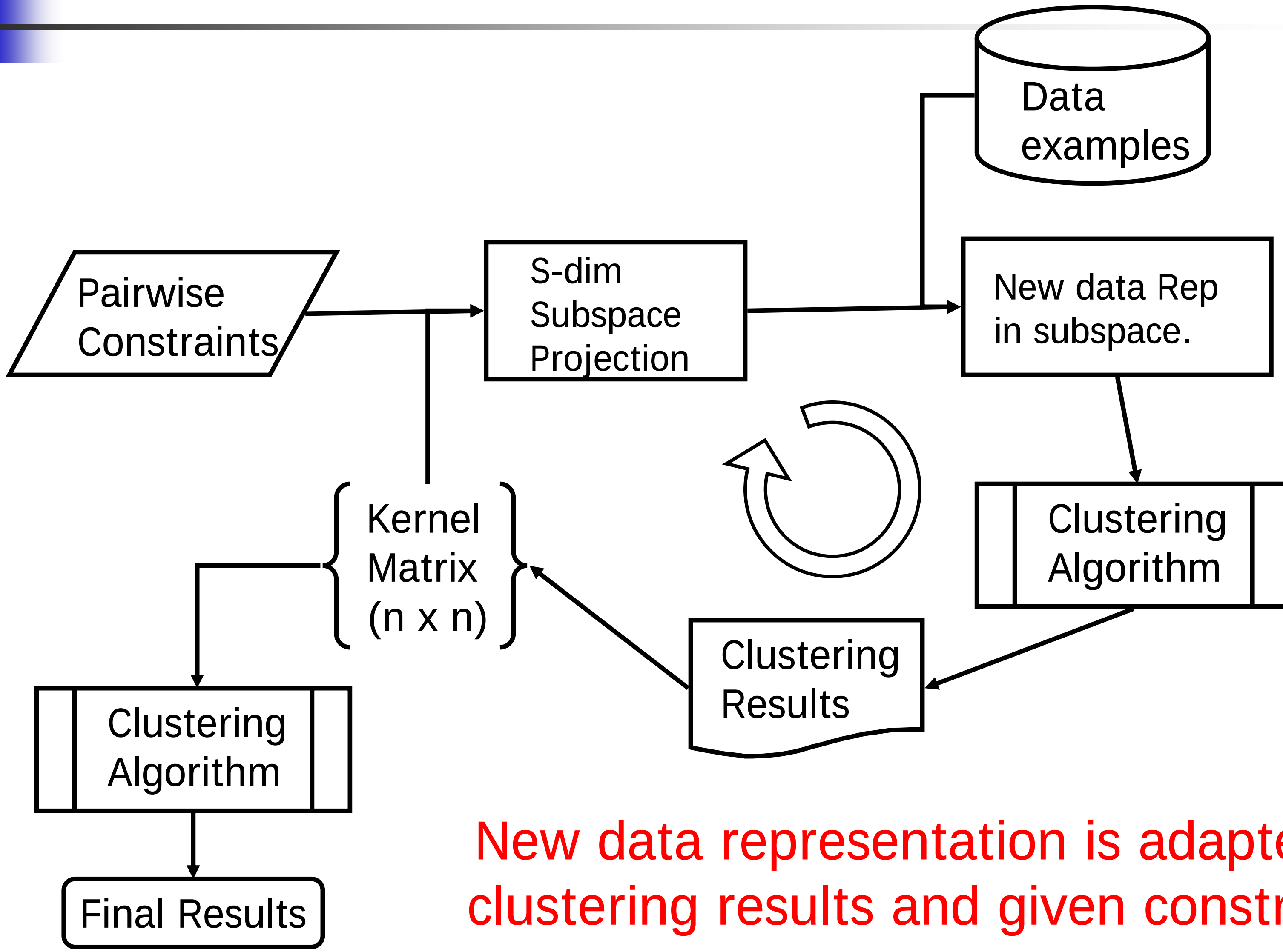
50% Training, 50% Testing; 5 labeled samples/class. Base classifier: SVM



BoostCluster

- A framework to improve any given clustering algorithm using pairwise constraints
- Basic Idea: Find a new data representation that encodes
  - the pairwise constraint information
  - the behavior of the underlying clustering algorithm.
- Boosting framework – “BoostCluster”

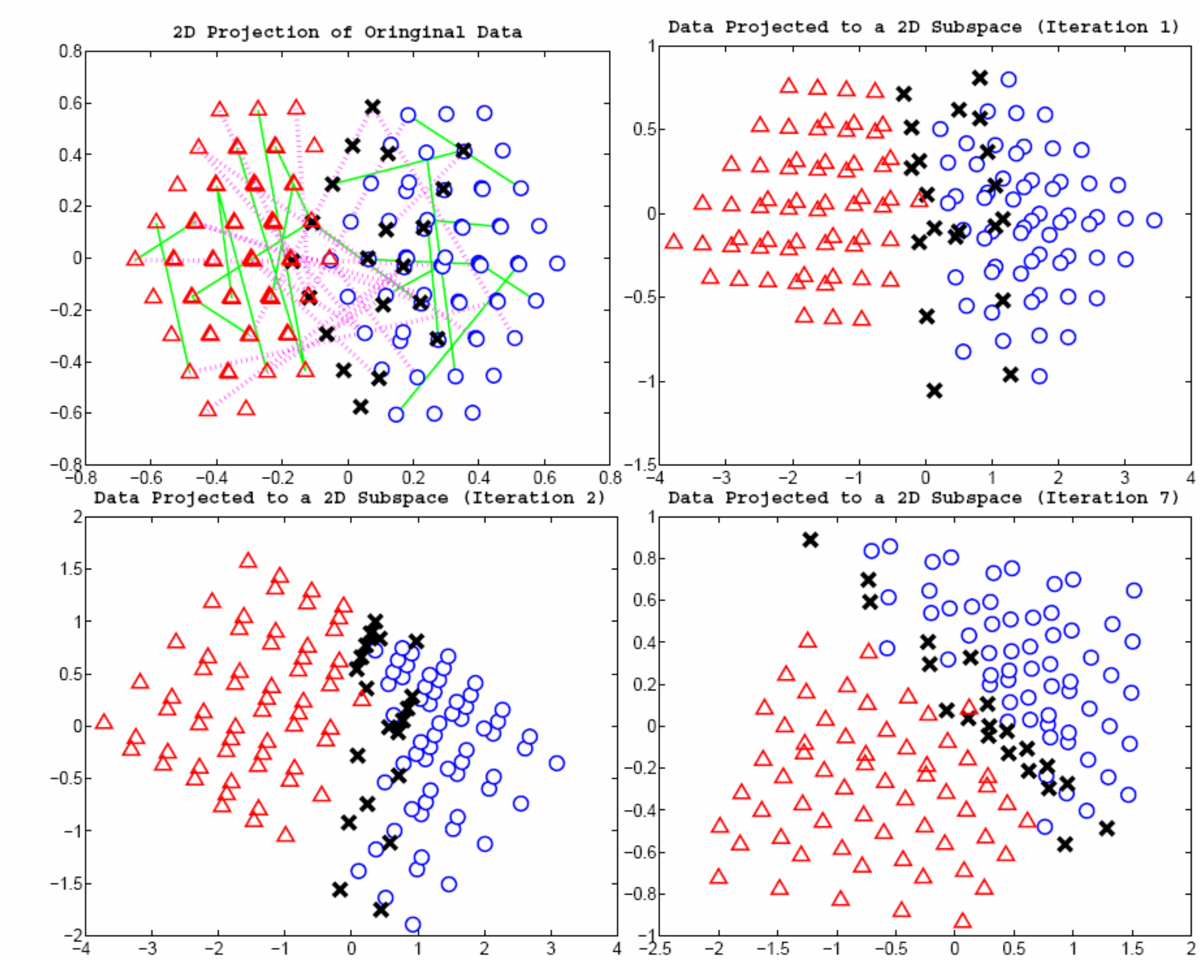
Boost Cluster



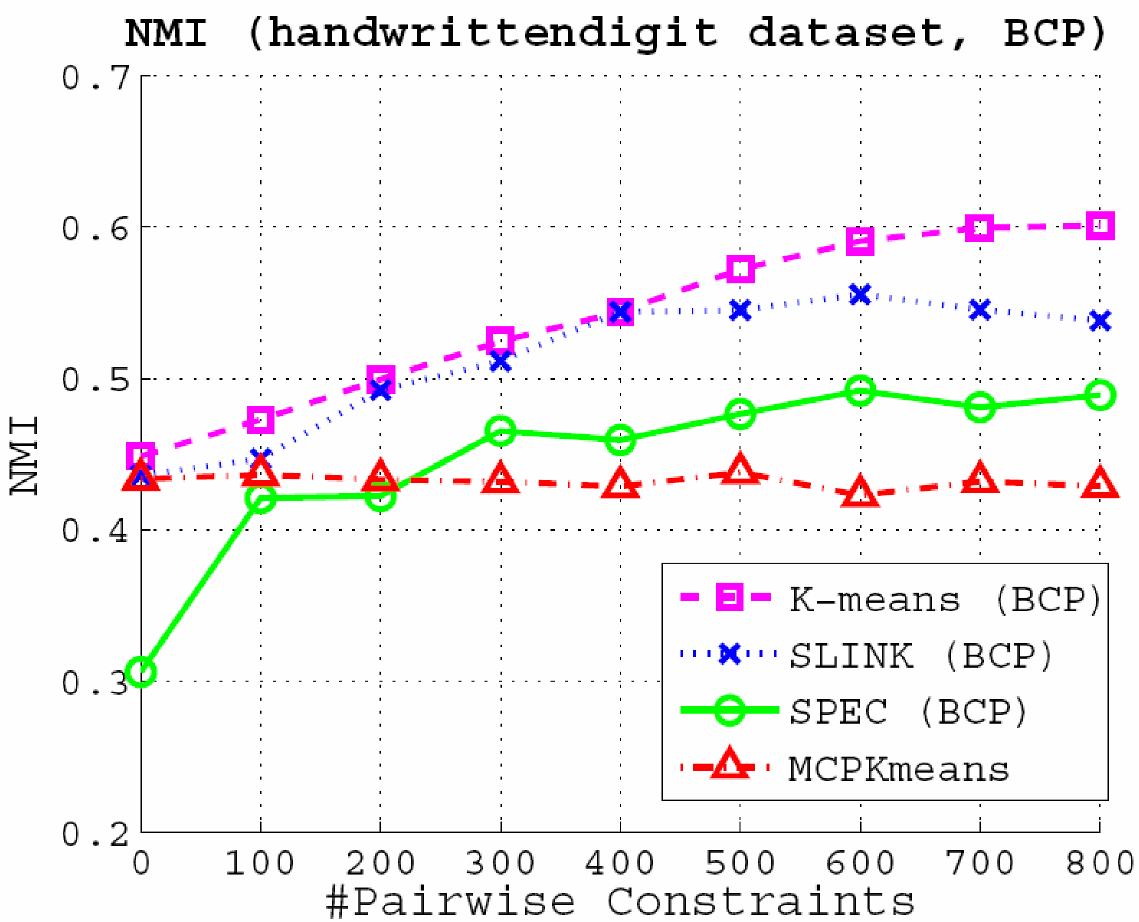
New data representation is adapted to clustering results and given constraints

Example

"Scale" data; 625 samples, 4 dimensions and 3 clusters

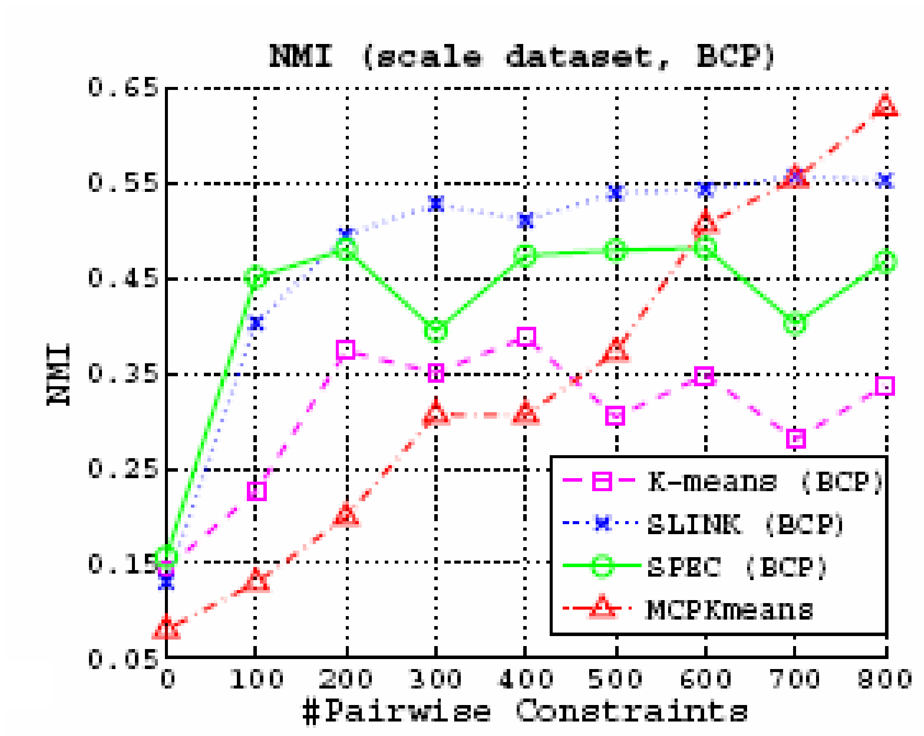


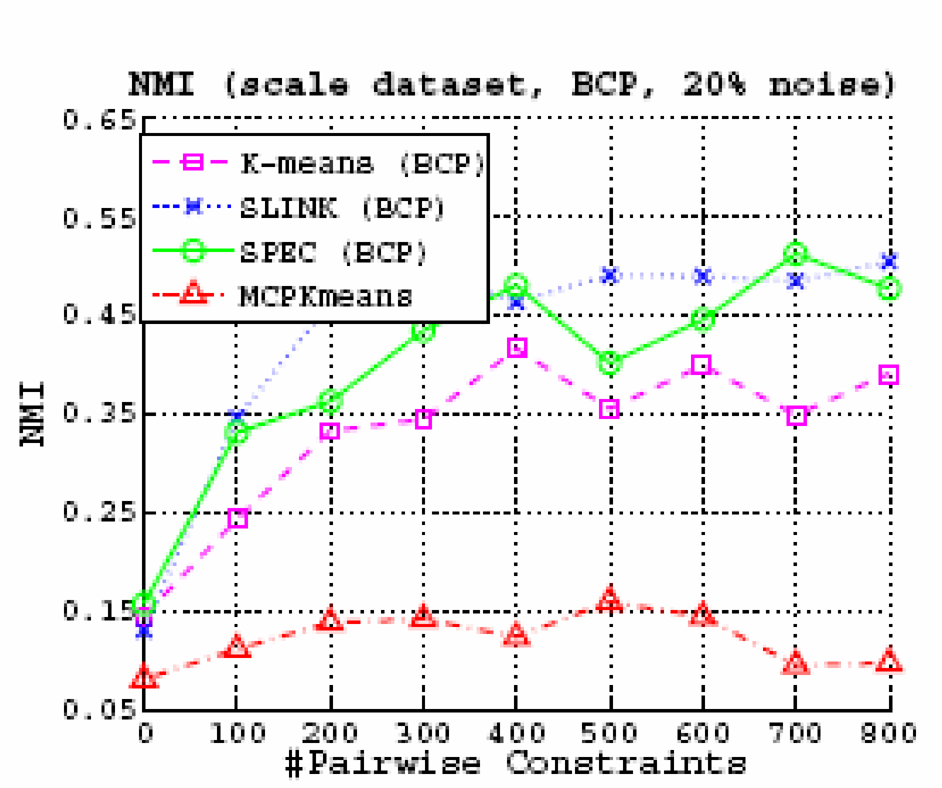
BoostCluster Performance



Basu, Bilenko and Mooney, A probabilistic framework for semi-supervised clustering, SIGKDD'04

BoostCluster Performance





Performance under noisy constraints (flip labels of 20% of randomly selected constraints)

Conclusions

- Semi-supervised learning is useful in situations where large amounts of data is readily available, but labeling the data is difficult
- Boosting-based framework is used to improve the performance of a classifier or clustering algorithm
- Experimental results show good performance improvement for a large variety of datasets
- Challenges: multiclass extension of SemiBoost; estimate the no. of clusters