ASCENT: Active Supervision for Semi-supervised Learning

TKDE-2019

Outline

- Method
  - Active Selection
  - Batch Mode Active Learning
- Experiments

Motivation

Active Selection:

Learning by Association : A versatile semi-supervised training method for neural networks **CVPR2017**

Batch Mode Active Learning:

Submodularity in Data Subset Selection and Active Learning **ICML2015**

Active Selection

Uncertainty :

A general assumption ：Good embeddings will have a high similarity if they belong to the same class

The uncertainty is measured on the similarity between labeled and unlabeled examples.

Definition 1: The similarity between embeddings A and B is defined as:

$$M_{ij} := A_i \cdot B_j, \tag{1}$$

Active Selection

Uncertainty:

Definition 2: The probabilities from A to B by softmaxing M over columns is

$$\begin{aligned} P_{ij}^{ab} &= P(B_j|A_i) := (softmax_{cols}(M))_{ij}, \\ &= exp(M_{ij}) / \sum_{j'} exp(M_{ij'}). \end{aligned} \tag{2}$$

Thus, the association probability of starting at A_i and ending up at A_j is

$$\begin{aligned} P_{ij}^{aba} &= (P^{ab} P^{ba})_{ij}, \\ &= \sum_{k} P_{ik}^{ab} P_{kj}^{ba}. \end{aligned} \tag{3}$$

Active Selection

Uncertainty:

The loss penalizes incorrect associations and encourages to correct the distribution.

$$\ell_{uncertainty} = L(T, P^{aba}), \tag{4}$$

with the uniform target distribution

$$T_{ij} = \begin{cases} 1/|class(A_i)| & class(A_i) = class(A_j), \\ 0 & otherwise, \end{cases} \tag{5}$$

Active Selection

Influence (Representativeness) :

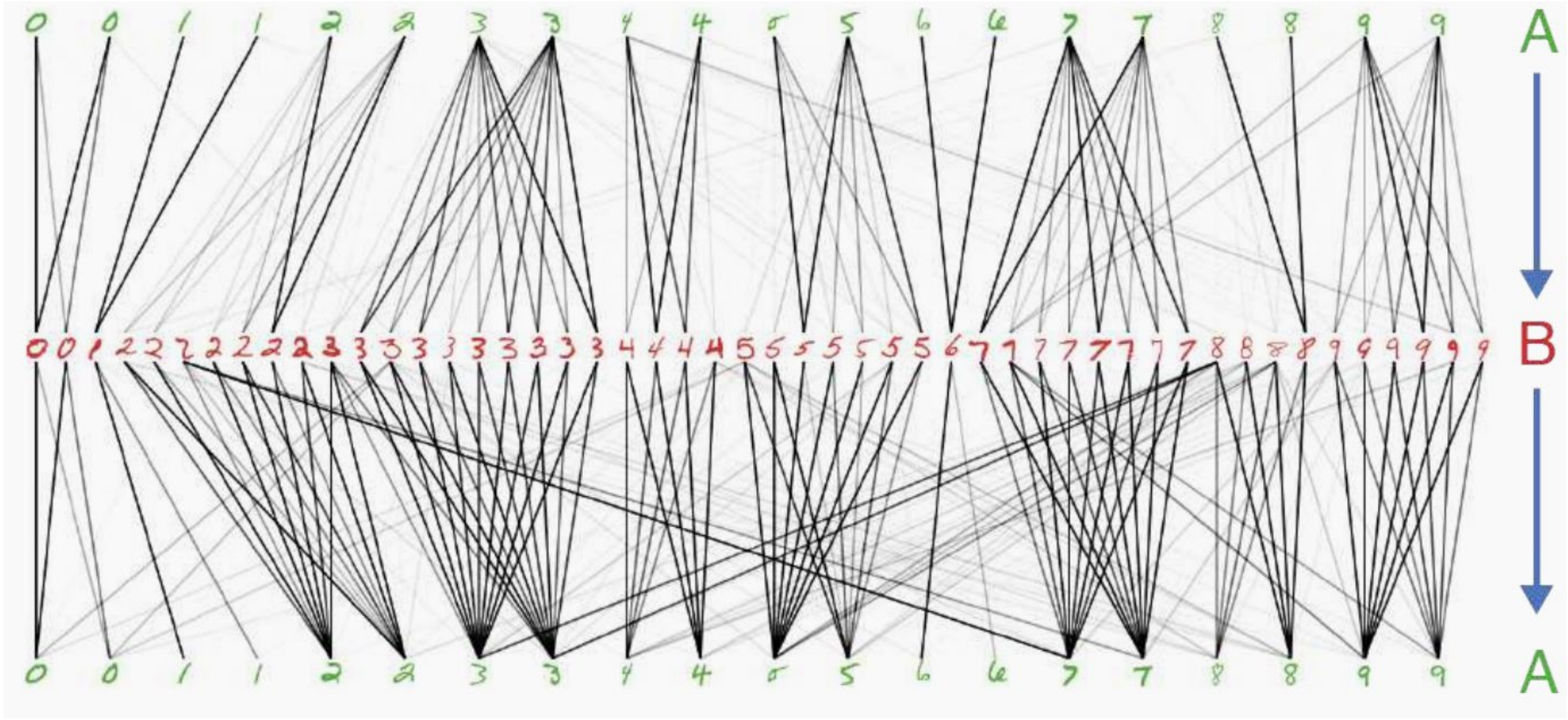
$$\ell_{influence} = L(V, P^{inf}), \tag{6}$$

where the influence probability of examples in B is $P_j^{inf} = P_{ij}^{ab}$ and the uniform target distribution is defined as $V_j = 1/|B|$.

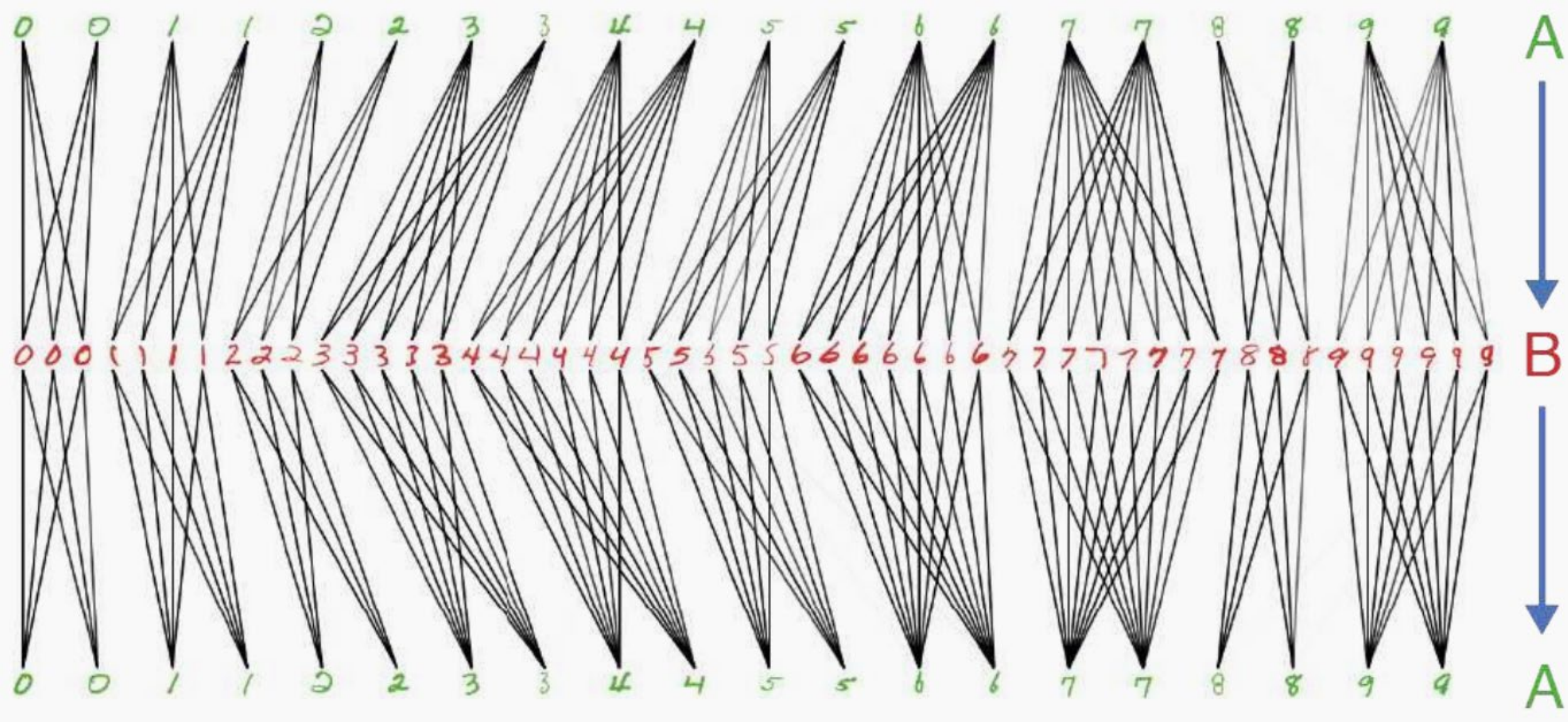
Dynamic combination:

$$\ell_{supervision} = \delta \ell_{uncertainty} + (1 - \delta)\ell_{influence}. \tag{7}$$

Active Selection Example



Active Selection Example



Batch Mode Active Learning

Which Examples to Query?

Choose from a batch of unlabeled examples U, which has the top **βt** highest score.

$$\text{that } m_y(\mathcal{S}) = \mathcal{S} \cap \mathcal{N}^y = K \frac{|\mathcal{N}^y|}{|\mathcal{N}|} \text{ for all } y \in \mathcal{Y} \text{ after balancing enforced and } |\mathcal{S}| = K.$$

Note: Use the most probable prediction yˆ according to the association probability

Batch Mode Active Learning

How to Query?

Use submodular functions to measure the utility of each subset

次模函数: 是一个集合函数，随着输入集合中元素的增加，增加单个元素到输入集合导致的函数增量的差异减小。

如果 Ω是一个有限集，一个子模函数是一个集函数 $f: 2^{\Omega} \to \mathbb{R}$，哪里 2^{Ω}表示功率设定的Ω，满足以下等效条件之一。

1. 每一个 $X, Y \subseteq \Omega$ 同 $X \subseteq Y$ 每一个 $x \in \Omega \setminus Y$ 我们有 $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$。
2. 每一个 $S, T \subseteq \Omega$ 我们有 $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$。

Batch Mode Active Learning

Define the submodular function

$$f(\mathcal{S}) = \sum_{i \in \mathcal{N}} \log p(\mathbf{x}^i | y^i; \theta(\mathcal{S})) + \sum_{i \in \mathcal{N}} \log p(y^i; \theta(\mathcal{S})),$$

Generative likelihood

$$P\left(x^i \middle| y^i; \theta(S)\right) = ce^{\|x^i - x^j\|_2^2} = ce^{w(i,j)-d} = c`e^{w(i,j)} = c`\exp(\max_{s \in S \cap N^{y^i}} w(i,s))$$

Prior likelihood

$$P\left(y^i; \theta(S)\right) = m_{y^i}(S)/|S| \quad \text{其中，} \ m_y(S) = \sum_{i \in N} 1\{\mathrm{y^i} = \mathrm{y}\}$$

Batch Mode Active Learning

$$f(\mathcal{S}) = \underbrace{\sum_{y\in\mathcal{Y}}\sum_{i\in\mathcal{N}^y}\max_{s\in\mathcal{S}\cap\mathcal{N}^{y^i}} w(i,s)}_{term\ 1: f(\mathcal{S})} + \underbrace{\sum_{y\in\mathcal{Y}} m_y(\mathcal{N})\log m_y(\mathcal{S})}_{term\ 2} - \underbrace{|\mathcal{N}|\log|\mathcal{S}|}_{term\ 3} + \underbrace{C}_{constant} . \quad F(S) \tag{9}$$

The first term is our submodular function:

$$f(\mathcal{S}) = \sum_{y\in\mathcal{Y}}\sum_{i\in\mathcal{N}^y}\max_{s\in\mathcal{S}\cap\mathcal{N}^{y^i}} w(i,s). \tag{10}$$

$$\max_{|\mathcal{S}|=K;\mathcal{S}\subseteq\mathcal{U}^t} f(\mathcal{S}). \tag{12}$$

Algorithm

Algorithm 1 Batch Mode Active Learning (ASCENT)

Input: $\mathcal{U}, T, K, \{\beta_t\}_{t=1}^{T}$, starting set of labels $\mathcal{L}$
repeat
  Training the classifier using the labeled set $\mathcal{L}$, and derive the supervision score ρ^t;
  $\mathcal{U}^t \in \arg\max_{u \subseteq \mathcal{U} \setminus \mathcal{L}; |\mathcal{U}| = \beta_t} \sum_{u \in \mathcal{U}} \rho_u^t$;
  Obtain the most probable labels as the hypothesized labels $\{\hat{y}_u\}_{u \in \mathcal{U}^t}$ and ground set $\mathcal{U}^t$;
  Instantiate $\hat{f}_t : 2^{\mathcal{U}^t} \to \mathbb{R}_+$ on the hypothesized label $\{\hat{y}_u\}_{u \in \mathcal{U}^t}$ and ground set $\mathcal{U}^t$;
  Find $\mathcal{L}^t \in \arg\max_{|\mathcal{S}| = K; \mathcal{S} \subseteq \mathcal{U}^t \setminus \mathcal{L}} f(\mathcal{S})$.
  $\mathcal{L} = \mathcal{L} \cup \mathcal{L}^t$
until $t > T$

some hyperparameters

- **δ** which trades off the uncertainty and influence of examples
- **{βt}** the size of candidate set U_t, which trades off between the criteria of supervision score and submodular objective f

Algorithm for Classification and Clustering

Algorithm 2 ASCENT for Semi-supervised Classification

Input: The selected optimal subset at each time v; The hyperparameters β; The mode M; The pool of unlabeled examples $\mathcal{D}^{\mathcal{U}}$; The initial training set $\mathcal{D}^{\mathcal{L}}$.
repeat
  $\mathcal{V} \leftarrow \text{ASCENT}(\mathcal{D}^{\mathcal{U}}, v, \mathcal{D}^{\mathcal{L}}, \beta)$
  $\mathcal{D}^{L} \leftarrow \mathcal{D}^{L} \cup \mathcal{V}$
  $\mathcal{D}^{U} \leftarrow \mathcal{D}^{U} \setminus \mathcal{V}$
  $M \leftarrow M(\mathcal{D}^{L})$
until achieve the budget
return M.

Algorithm 3 ASCENT for Semi-supervised Clustering

Input: A set of data points $\mathcal{D}$; the total number of classes c; The selected optimal subset at each time v; The hyperparameters β.
Initializations: $\mathcal{C} = \emptyset$; $\Sigma = \{\Sigma_1, ..., \Sigma_k\}$
repeat
  π=SSC($\mathcal{D}$,$\mathcal{C}$);
  $\mathcal{V}$=ASCENT(π, Σ, v, β)
  for each $\mathbf{x}_i$ in $\mathcal{V}$ **do**
    for each $\Sigma_j \in \Sigma$ in decreasing order of probability $P(\mathbf{x}_i \in \Sigma_j)$ **do**
      Query $\mathbf{x}_i$ against any data points $\mathbf{x}_k \in \Sigma_j$;
      Update $\mathcal{C}$ based on returned answer;
      if $(\mathbf{x}_i, \mathbf{x}_k, M)$ **then**
        $\Sigma_j = \Sigma_j \cup \{\mathbf{x}_i\}$;break;
      end if
    end for
    if no must-link is achieved **then**
      $l++; \Sigma_l = \mathbf{x}_i; \Sigma = \Sigma \cup \Sigma_l$;
    end if
  end for
until achieve the budget
return SSC($\mathcal{D}$,$\mathcal{C}$)

Experiments

Datasets:

- MNIST
- CIFAR-10^3
- IMDB
- RCV1

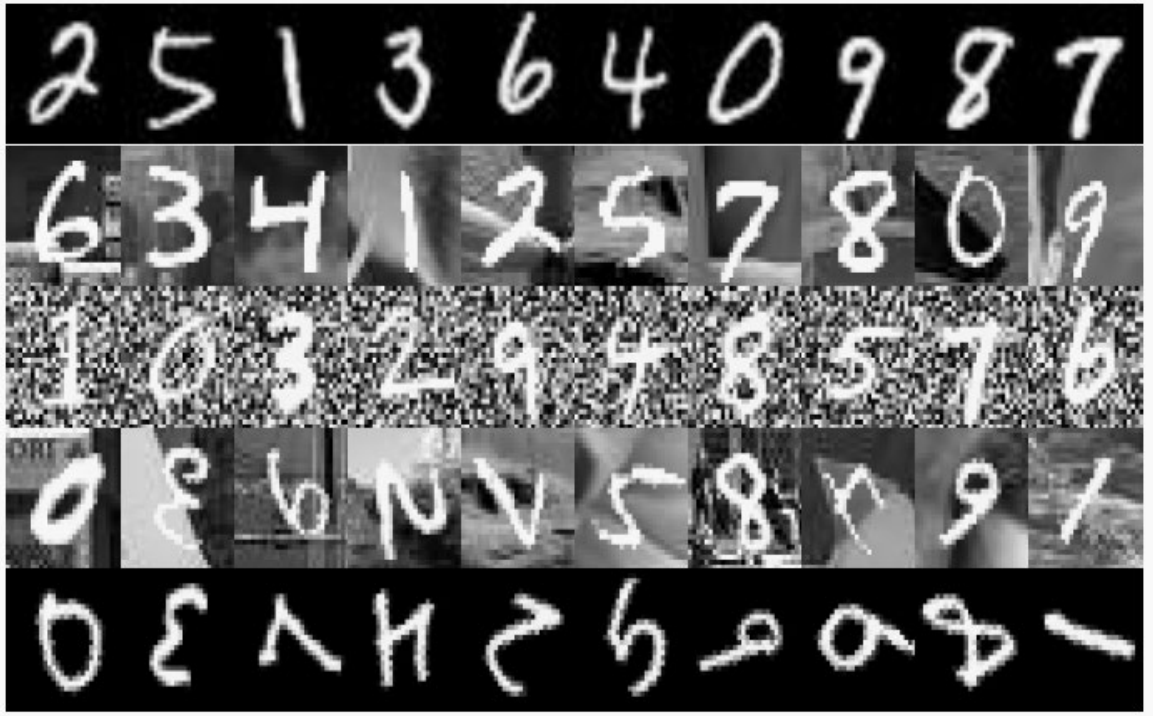


Fig. 1. Samples from MNIST dataset and its variations

Classification Experiments

Datasets	ASCENT vs. Baselines (ConvNet)					
	RAND	S-MARGIN	QUIRE	MMC	FASS	EE-BMAL
M_{basic}	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$
M_{rot}	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$
M_{rand}	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$
M_{img}	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$
M_{rotImg}	$< .05$	$< .05$	$< .05$	$< .1$	$< .1$	$< .1$
CIFAR-10	$< .05$	$< .05$	$< .05$	$< .1$	$< .05$	$< .1$
IMDB	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$
RCV1	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$

ASCENT vs. Baselines (Best $k-$NN)					
RAND	S-MARGIN	QUIRE	MMC	FASS	EE-BMAL
$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$
$< .05$	$< .05$	$< .05$	$< .1$	$< .05$	$< .1$
$< .05$	$< .05$	$< .05$	$< .1$	$< .1$	$< .05$
$< .05$	$< .05$	$< .05$	$< .05$	$< .1$	$< .1$
$< .1$	$< .1$	$< .1$	$< .1$	$< .1$	$< .05$
$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$
$< .05$	$< .05$	$< .05$	$< .1$	$< .05$	$< .1$
$< .05$	$< .05$	$< .05$	$< .05$	$< .05$	$< .05$

performs significantly better ($p < .05$) than the compared approaches

Classification Experiments

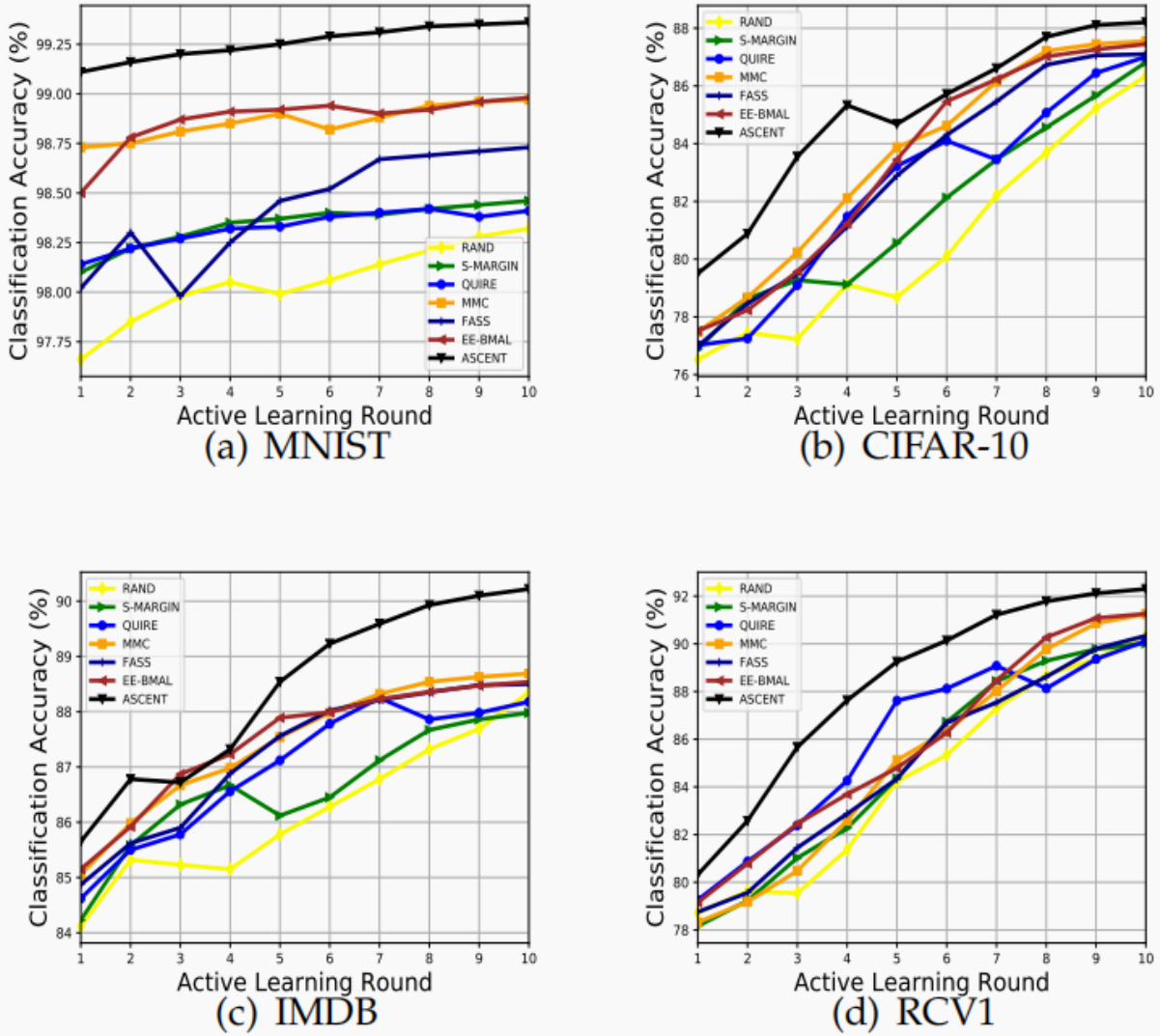


Fig. 2. Comparison results of different active learning algorithms by taking ConvNets as the base classification model on the datasets

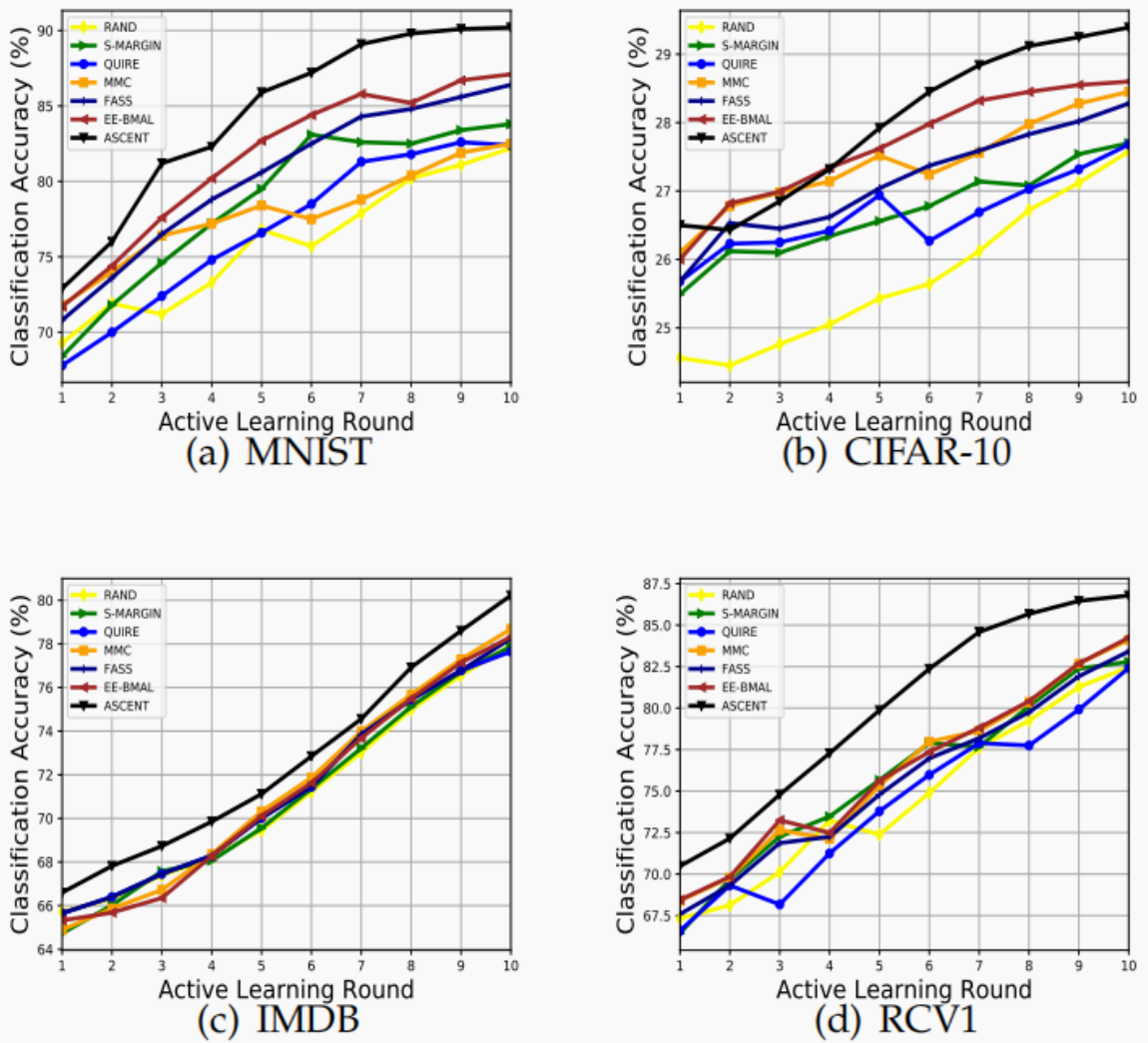


Fig. 3. Comparison results of different active learning algorithms by taking best k−NN as the base classification model on the datasets

Classification Experiments

More Analysis on SSL Methods

- **label propagation** [59]
- **learning by association** [21]

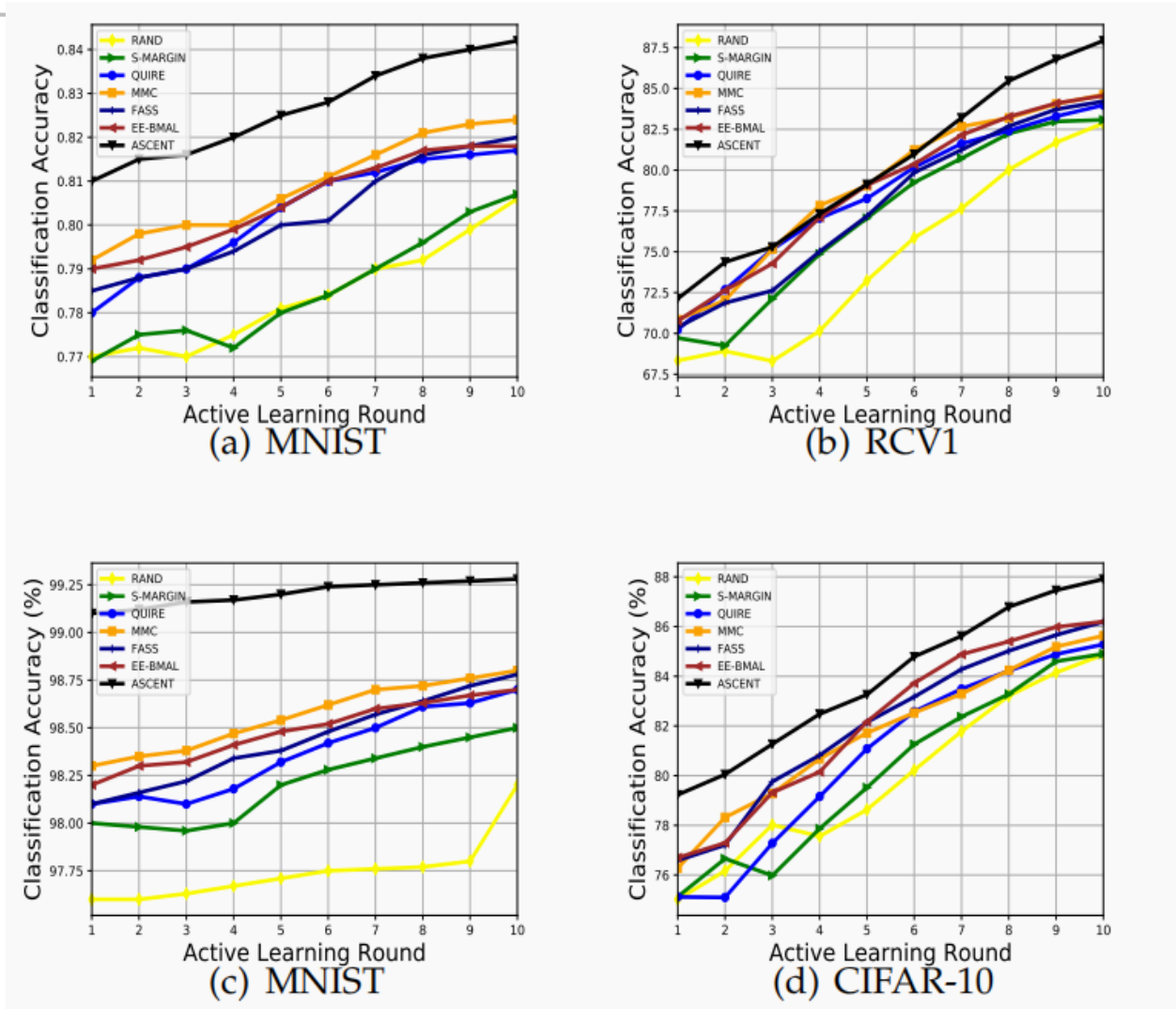
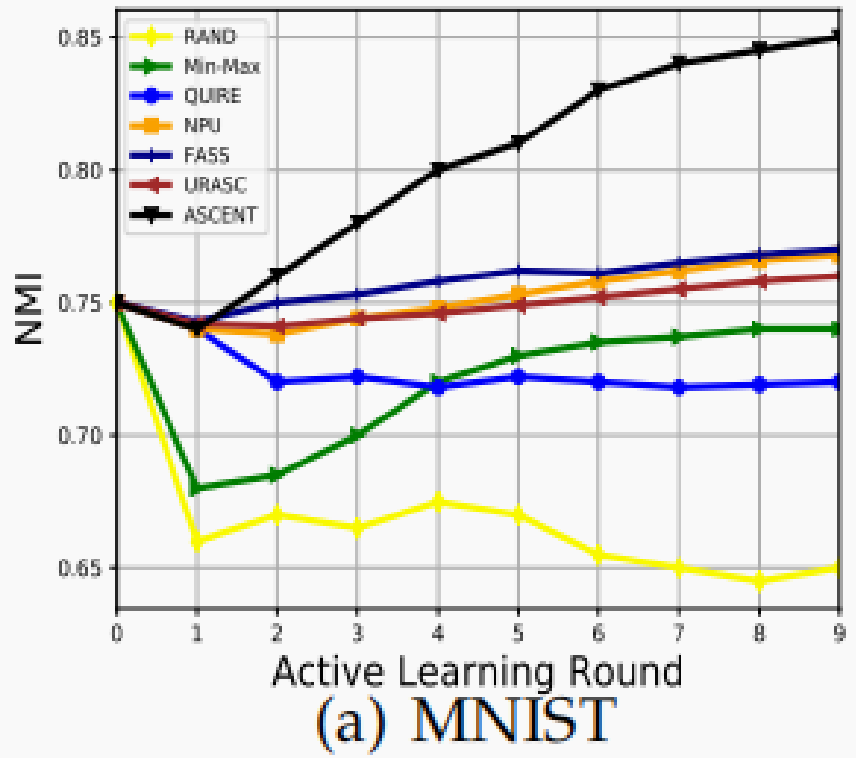


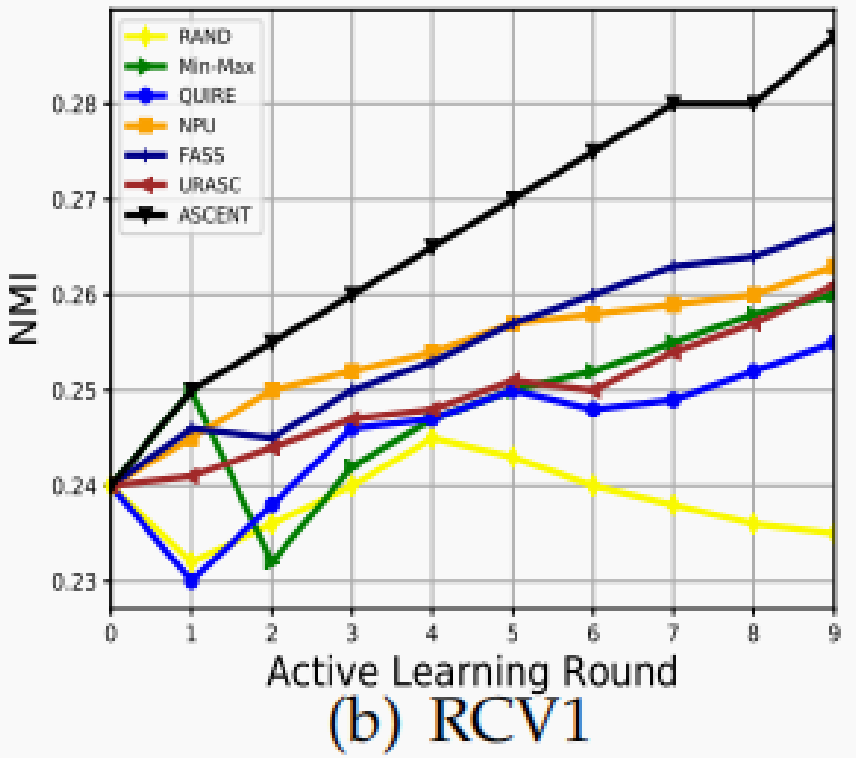
Fig. 4. Comparison results of different active learning algorithms by taking label propagation (a-b) and learning by association (c-d) as the base classification models on the datasets

Clustering Experiments

$$NMI = \frac{2I(c;k)}{H(c) + H(k)},$$



(a) MNIST



(b) RCV1

Fig. 5. The NMI of different methods on the datasets

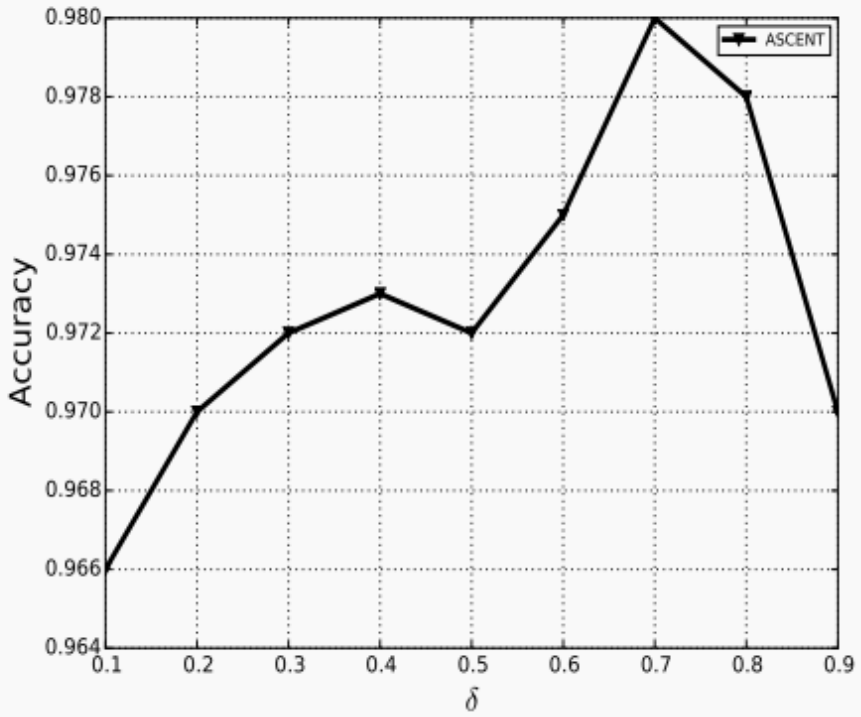
Clustering Experiments

Datasets	Methods	Number of queries								
		30	45	60	75	90	105	120	135	150
MNIST	RAND	0.76±0.10	0.77±0.15	0.77±0.10	0.76±0.10	0.75±0.10	0.75±0.13	0.75±0.11	0.75±0.10	0.75±0.11
	Min-Max	0.81±0.08	0.80±0.10	0.79±0.09	0.79±0.07	0.80±0.09	0.82±0.06	0.84±0.00	0.84±0.00	0.84±0.00
	QUIRE	0.77±0.10	0.78±0.08	0.78±0.06	0.80±0.06	0.79±0.01	0.81±0.10	0.82±0.01	0.83±0.11	0.83±0.02
	NPU	0.81±0.07	0.82±0.05	0.83±0.06	0.83±0.03	0.84±0.03	0.84±0.01	0.85±0.00	0.85±0.00	0.85±0.00
	FASS	0.81±0.09	0.83±0.17	0.86±0.01	0.86±0.03	0.87±0.01	0.87±0.02	0.87±0.03	0.89±0.01	0.89±0.02
	URASC	0.81±0.08	0.83±0.15	0.85±0.01	0.86±0.05	0.86±0.01	0.87±0.03	0.87±0.02	0.88±0.05	0.88±0.03
	ASCENT	**0.82±0.10**	**0.85±0.01**	**0.87±0.01**	**0.88±0.03**	**0.88±0.02**	**0.89±0.02**	**0.89±0.01**	**0.90±0.01**	**0.91±0.01**
		200	300	400	500	600	700	800	900	1000
RCV1	RAND	0.13±0.02	0.13±0.01	0.13±0.02	0.13±0.01	0.14±0.02	0.14±0.01	0.14±0.02	0.14±0.02	0.14±0.01
	Min-Max	0.15±0.04	0.15±0.02	0.15±0.01	0.15±0.00	0.15±0.02	0.15±0.01	0.15±0.00	0.15±0.01	0.15±0.00
	QUIRE	0.14±0.03	0.14±0.02	0.14±0.01	0.14±0.00	0.14±0.01	0.14±0.00	0.13±0.02	0.14±0.01	0.14±0.01
	NPU	0.16±0.02	**0.17±0.02**	0.16±0.00	0.16±0.01	0.16±0.01	0.16±0.00	0.17±0.01	0.17±0.00	0.17±0.00
	FASS	0.16±0.01	0.16±0.01	0.16±0.01	**0.17±0.02**	0.16±0.02	0.16±0.02	0.17±0.02	0.17±0.01	0.17±0.01
	URASC	0.16±0.01	0.16±0.02	0.16±0.01	0.16±0.00	0.15±0.01	0.15±0.02	0.16±0.00	0.16±0.00	0.16±0.00
	ASCENT	0.16±0.03	0.16±0.02	0.16±0.01	**0.17±0.02**	0.17±0.02	0.17±0.01	**0.18±0.01**	**0.18±0.01**	**0.18±0.01**

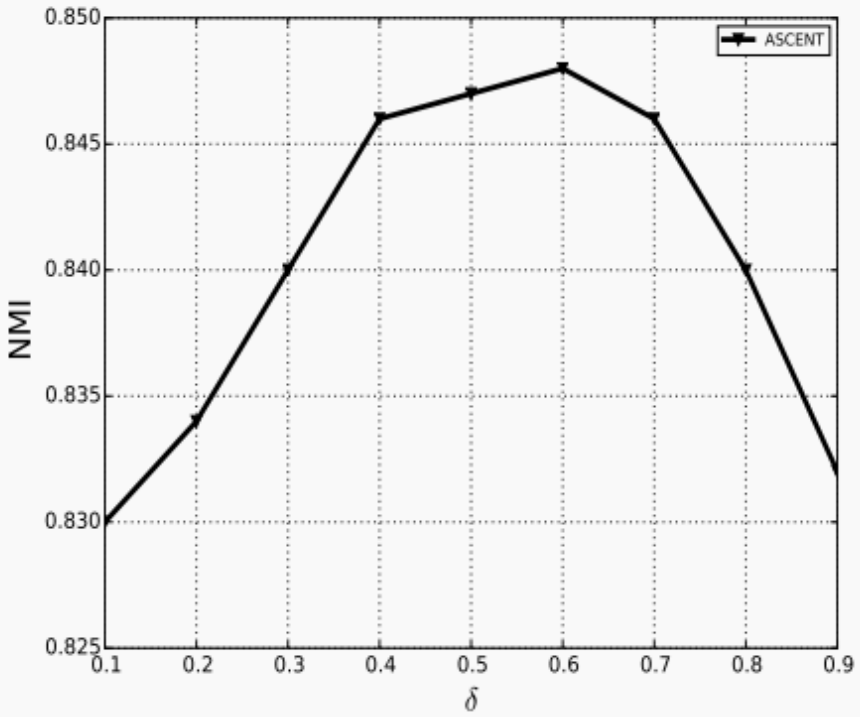
$$F - measure = \frac{2 \times Prediction \times Recall}{Prediction + Recall}$$

Experiments

Parameters Study



(a) Classification

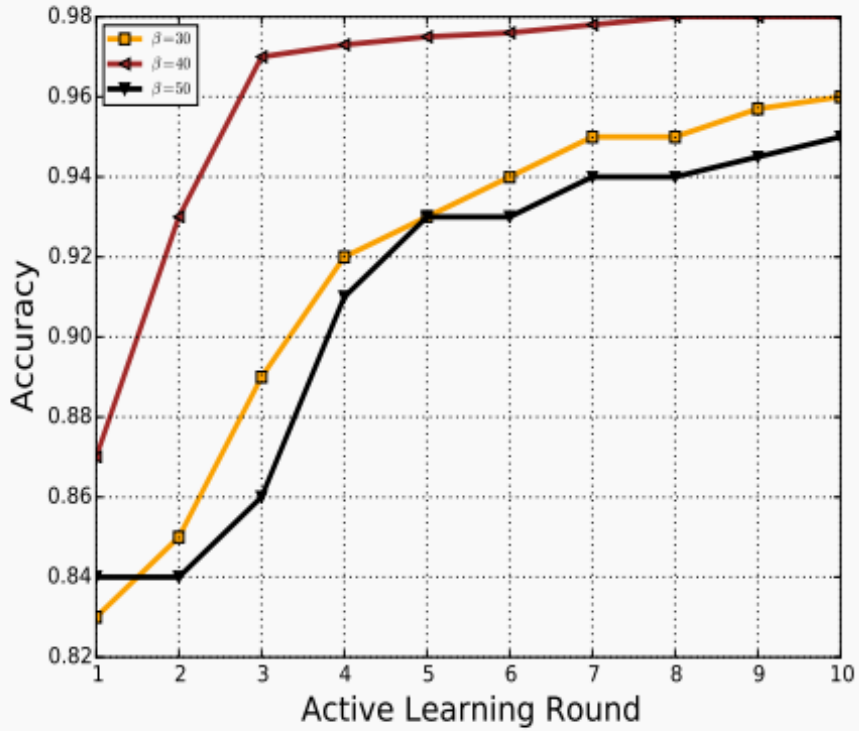


(b) Clustering

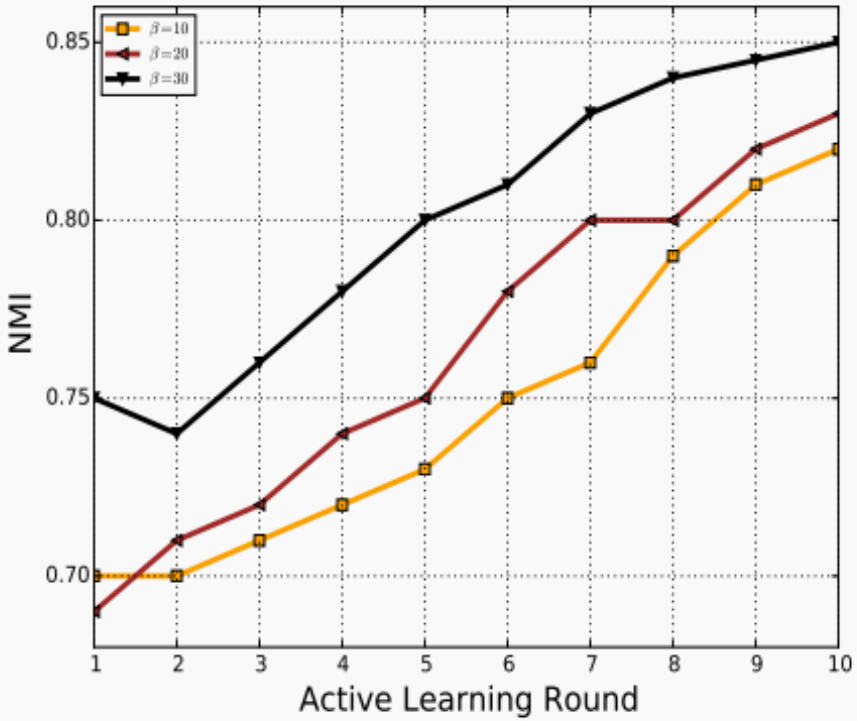
Fig. 11. The effect of δ

Experiments

Parameters Study



(a) Classification



(b) Clustering

Fig. 12. The effect of β