电子科技大学
University of Electronic Science and Technology of China

The Instance Selection For Active Learning

Dongzi Chen

Data Mining Lab,
Big Data Research Center, UESTC
Email：chendz321@hotmail.com

Outline

- What is Active Learning
- Instance Selection for Active Learning
- Asymmetric Active Learning with Imbalanced Data
- One More Thing about Instance Selection

Exploiting unlabeled data

A lot of unlabeled data is plentiful and cheap, eg.

documents off the web

speech samples

images and video

But labeling can be expensive.

Typical heuristics for active learning

- Start with a pool of unlabeled data
- Pick a few points at random and get their labels
- Repeat

    Fit a classifier to the labels seen so far

    Choose the instances to query their unlabeled point

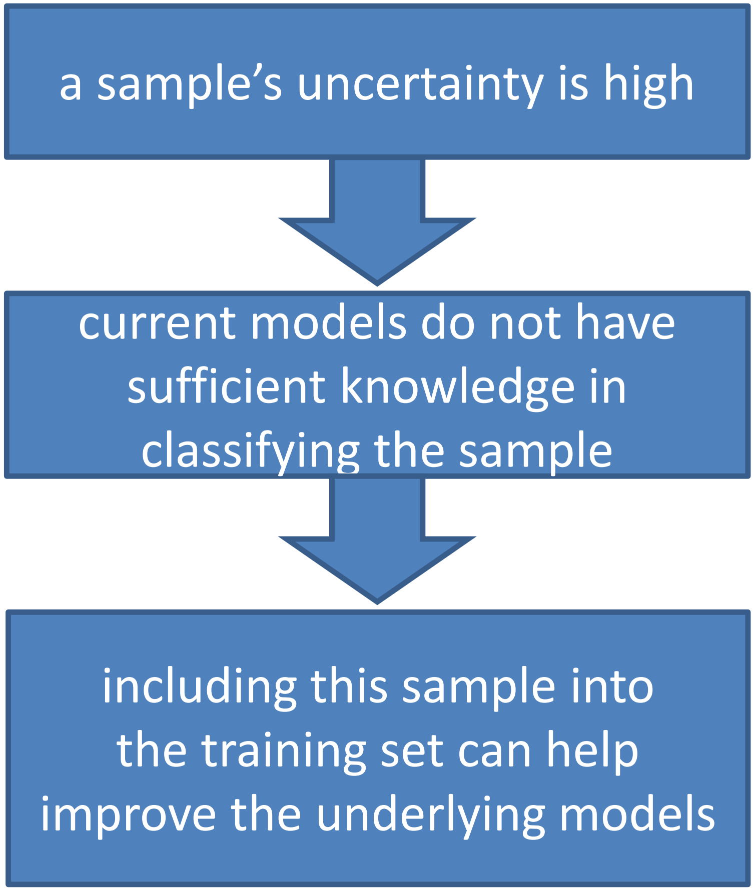
Key Point: How to choose instance?

Instance Selection for Active Learning

Most informative instances:

Uncertainty:

How to measure it ?



a sample's uncertainty is high

current models do not have sufficient knowledge in classifying the sample

including this sample into the training set can help improve the underlying models

Instance Selection for Active Learning

➤*Uncertainty sampling*

Least Confidence:

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}[1 - P_\Theta(\hat{y}|x)]$$

where $\hat{y}$ is the most likely class label with the highest posterior probability in the hypothesis

This method prefers the instances on which the current hypothesis has the least confidence in deciding their most likely class labels

➤*Uncertainty sampling*

Sample margin:

Margin approach is prone to selecting instances with minimum margin between posterior probabilities of the two most likely class labels, which is represented by

$$x_M^* = \underset{x}{\operatorname{argmin}}[P_\Theta(\hat{y}_1|x) - P_\Theta(\hat{y}_2|x)]$$

where$\hat{y}_1$and $\hat{y}_2$ are the first and second most probable class labels.

The most informative instances are the ones with the smallest margins between the top two class labels.

Instance Selection for Active Learning

➢*Uncertainty sampling*

Entropy:

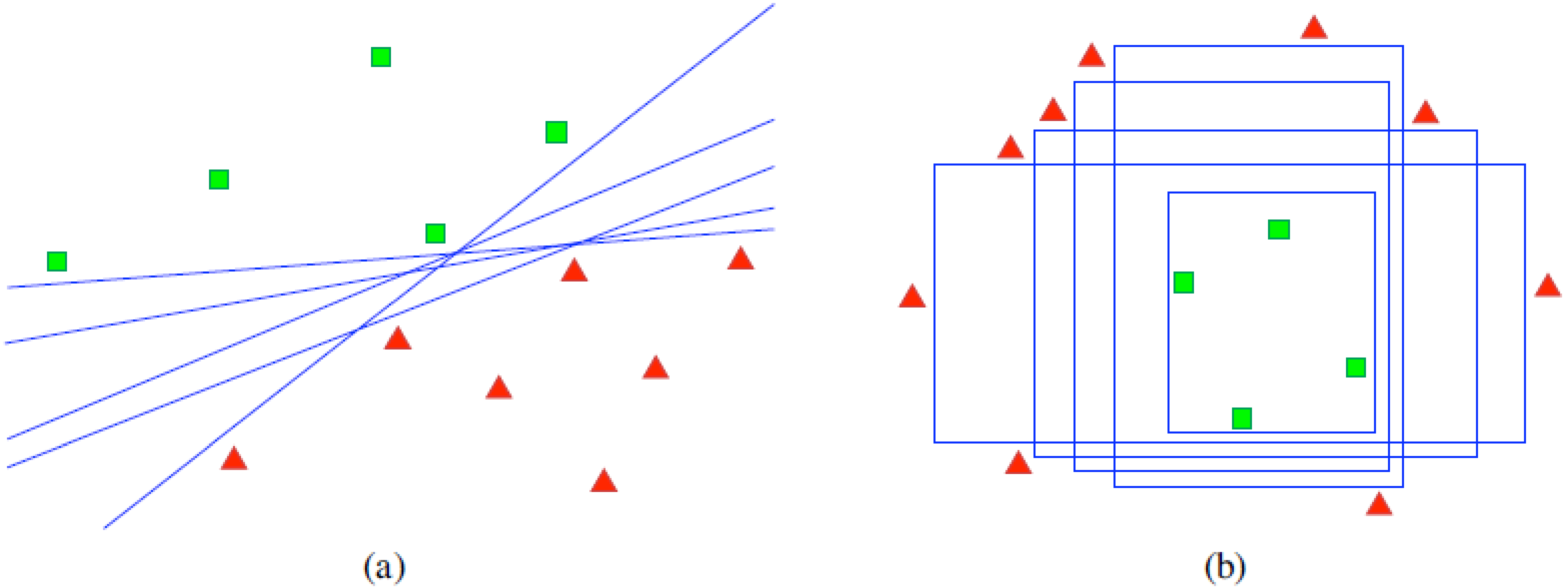
Given a hypothesis Θ, the prediction distribution of an instance x_k, then the uncertainty can be encoded as follows:

$$x_E^* = \underset{x}{\operatorname{argmax}} \sum_i P_\Theta(\hat{y}_i|x_k) log P_\Theta(\hat{y}_i|x_k)$$

where $\hat{y}_i$ denotes posterior probability of the instance x_k being a member of the i^{th} class, which ranges over all possible labels.

For a binary classification task, the most potential instances are the ones with equal posterior probability with respect to all possible classes.

➢Query-By-Committee



Version space examples for (a) linear and (b) axis-parallel box classifiers. All hypotheses are consistent with the labeled training data in L (as indicated by shaded polygons), but each represents a different model in the version space.

➢Query-By-Committee

Measuring the Level of Disagreement

Vote Entropy:

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} log \frac{V(y_i)}{C}$$

Average Kullback-Leibler (KL)divergence:

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{C} \sum_{c=1}^{C} D(P_{\theta^{(c)}} || P_{\mathrm{C}})$$

Where: $D(P_{\theta^{(c)}} || P_{\mathrm{C}}) = \sum_i P_{\theta^{(c)}}(y_i|x) log \frac{P_{\theta^{(c)}}(y_i|x)}{P_{\mathrm{C}}(y_i|x)}$

➤*Expected* Model Change

Idea: selecting the instance that would impart the greatest change to the current model if we knew its label.

Expected Gradient Length (EGL) approach

Let $\nabla l_\theta(L)$ be the gradient of the objective function l with respect to the model parameters θ.

Let $\nabla l_\theta(L \cup \langle x, y_i \rangle)$ be the new gradient that would be obtained by adding the training tuple $\langle x, y_i \rangle$ to L.

$$x^*_{EGL} = \underset{x}{\operatorname{argmax}} \sum_i P_\Theta(\hat{y}_i|x) \|\nabla l_\theta(L \cup \langle x, y_i \rangle)\|$$

where $\|\cdot\|$ is the Euclidean norm of each resulting gradient vector

➢Expected Error Reduction

Idea: to estimate the expected future error of a model trained using $L \cup \langle x, y_i \rangle$ on the remaining unlabeled instances in U and query the instance with minimal expected future error (sometimes called risk).

➤Expected Error Reduction

Minimal expected 0/1-loss:

$$x^*_{0/1} = \underset{x}{\operatorname{argmin}} \sum_i P_\Theta(\hat{y}_i|x)\left(\sum_{u=1}^{U} 1 - P_{\Theta^{+\langle x,y_i\rangle}}\left(\hat{y}_i\middle|x^{(u)}\right)\right)$$

Where: $\Theta^{+\langle x,y_i\rangle}$ refers to the new model after it has been re-trained with the training tuple $\langle x, y_i\rangle$ added to L

Instance Selection for Active Learning

➢Expected Error Reduction

Minimize the expected log-loss:

$$x^*_{log} = \underset{x}{\operatorname{argmin}} \sum_i P_\Theta(\hat{y}_i|x)\left(-\sum_{u=1}^{U}\sum_j P_{\Theta^{+\langle x,y_i\rangle}}\left(\hat{y}_j\middle|x^{(u)}\right)\log P_{\Theta^{+\langle x,y_i\rangle}}\left(\hat{y}_j\middle|x^{(u)}\right)\right)$$

Where: $\Theta^{+\langle x,y_i\rangle}$ refers to the new model after it has been re-trained with the training tuple $\langle x, y_i\rangle$ added to L

➢*Variance reduction*

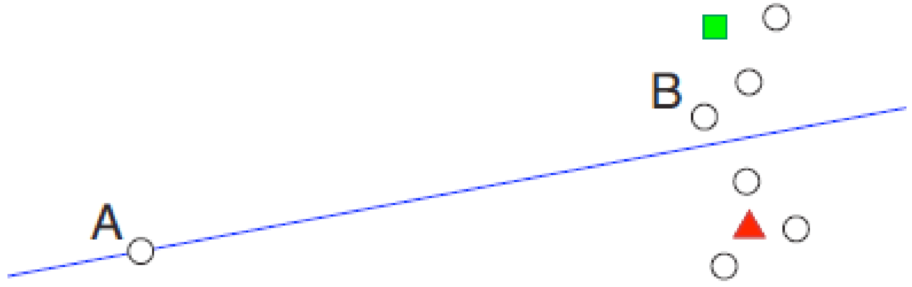
Minimizing the expectation of a loss function directly is expensive, and in general this cannot be done in closed form. However, we can still reduce generalization error indirectly by minimizing output variance, which sometimes does have a closed-form solution.

$$\text{Let } \sigma_{\hat{y}}^2 = E_L[(\hat{y} - E_L[\hat{y}])^2]$$

$$x_{VR}^* = \underset{x}{\operatorname{argmax}} \langle \sigma_{\hat{y}}^2 \rangle^{+x}$$

➢Density-Weighted Methods

An illustration of when uncertainty sampling can be a poor strategy for classification:



Shaded polygons represent labeled instances in L, and circles represent unlabeled instances in U. Since A is on the decision boundary, it would be queried as the most uncertain. However, querying B is likely to result in more information about the data distribution as a whole.

Instance Selection for Active Learning

➢Density-Weighted Methods

$$x_{ID}^{*} = \underset{x}{\operatorname{argmax}}\, \emptyset_A(x) \times \left(\frac{1}{U} \sum_{u=1}^{U} sim\left(x, x^{(u)}\right) \right)^{\beta}$$

$\emptyset_A(x)$ represents the informativeness of x according to some "base" query strategy A, such as an uncertainty sampling or QBC approach.

The second term weights the informativeness of x by its average similarity to all other instances in the input distribution (as approximated by U), subject to a parameter β that controls the relative importance of the density term.

Asymmetric Query Model

➢Binary Classification For Balanced Data

$$\Pr(Query(p_t) = 1) = \frac{c}{|p_t| + c}$$

Where: $p_t = {w_{t-1}}^T x_t$, c > 0 is a parameter that determines the threshold of uncertainty.

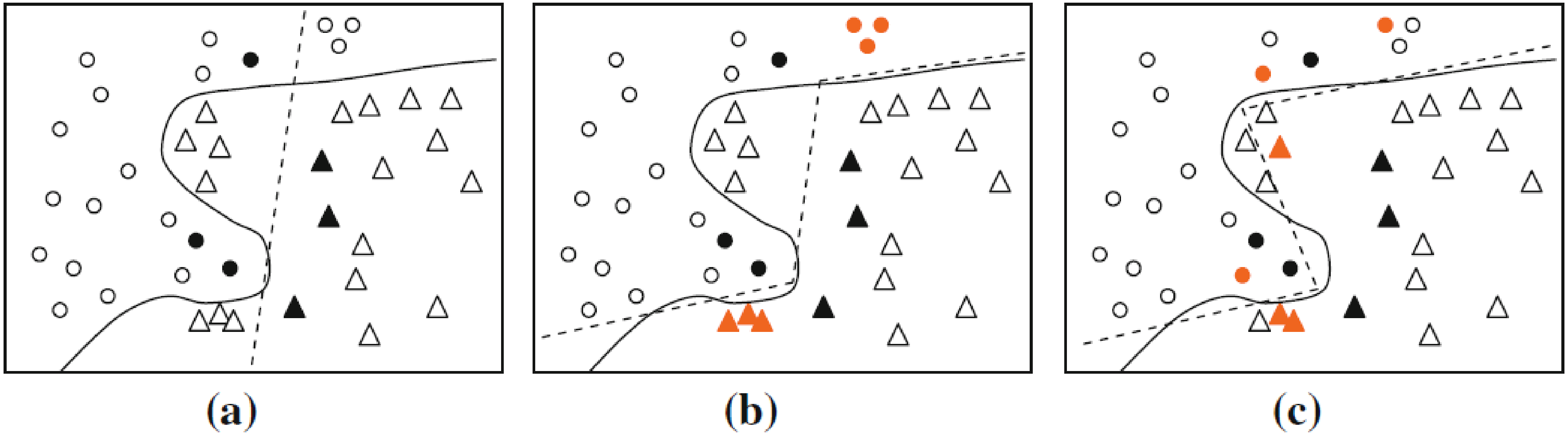
➢Binary Classification For Imbalanced Data

If there are many more negative examples than positive examples, treating them equally for making the query decisions can be sub-optimal.

$$\Pr(Query(p_t) = 1) = \begin{cases} \dfrac{c_+}{|p_t| + c_+} & if\ p_t \geq 0 \\ \dfrac{c_-}{|p_t| + c_-} & otherwise \end{cases}$$

If the negative class is the dominant class, then it is expected that c_- should be set to a larger value than c_+

Diversity



a Decision boundary learned from six labeled training instances.
b By labeling six most uncertain instances, the learner refines its decision boundary.
c By taking sample diversity into consideration, a method chooses the most informative candidate instances with low redundancy between them, based on which the learned decision boundary is significantly improved.

Thanks