A Cost-sensitive Active Learning for Imbalance Data with Uncertainty and Diversity Combination

Huailong Dong
School of Computer Science and Engineering,
Nanjing University of Science and Technology
200 Xiaolingwei Street, Nanjing 210094, China
dhl0010@163.com

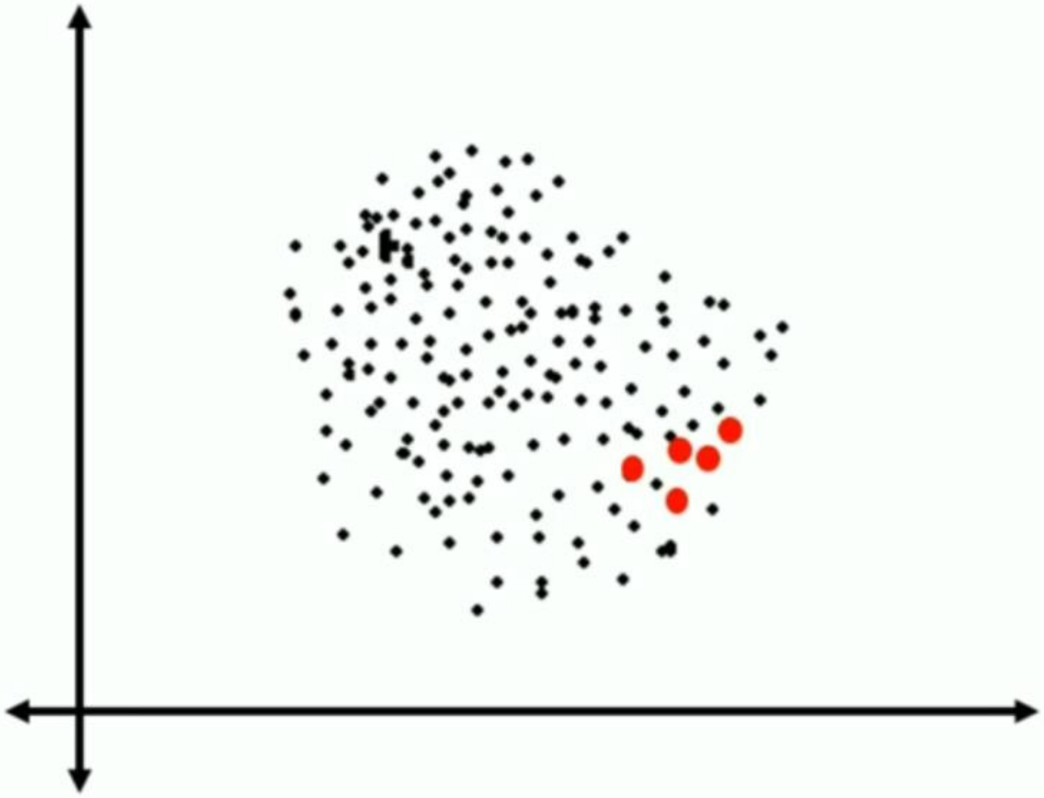
Bowen Zhu
School of Computer Science and Engineering,
Nanjing University of Science and Technology
200 Xiaolingwei Street, Nanjing 210094, China
2391986553@qq.com

Jing Zhang*
School of Computer Science and Engineering,
Nanjing University of Science and Technology
200 Xiaolingwei Street, Nanjing 210094, China
jzhang@njust.edu.cn

ICMLC 2020

Introduction

The class distributions of real-world classification datasets are usually imbalanced because many applications, such as network intrusion detection, tumor classification, financial risk identification, etc.



Two Class Classification

Two difficulties that we may face are：

imbalanced class distributions → Imbalanced learning

selecting proper training samples → active learning

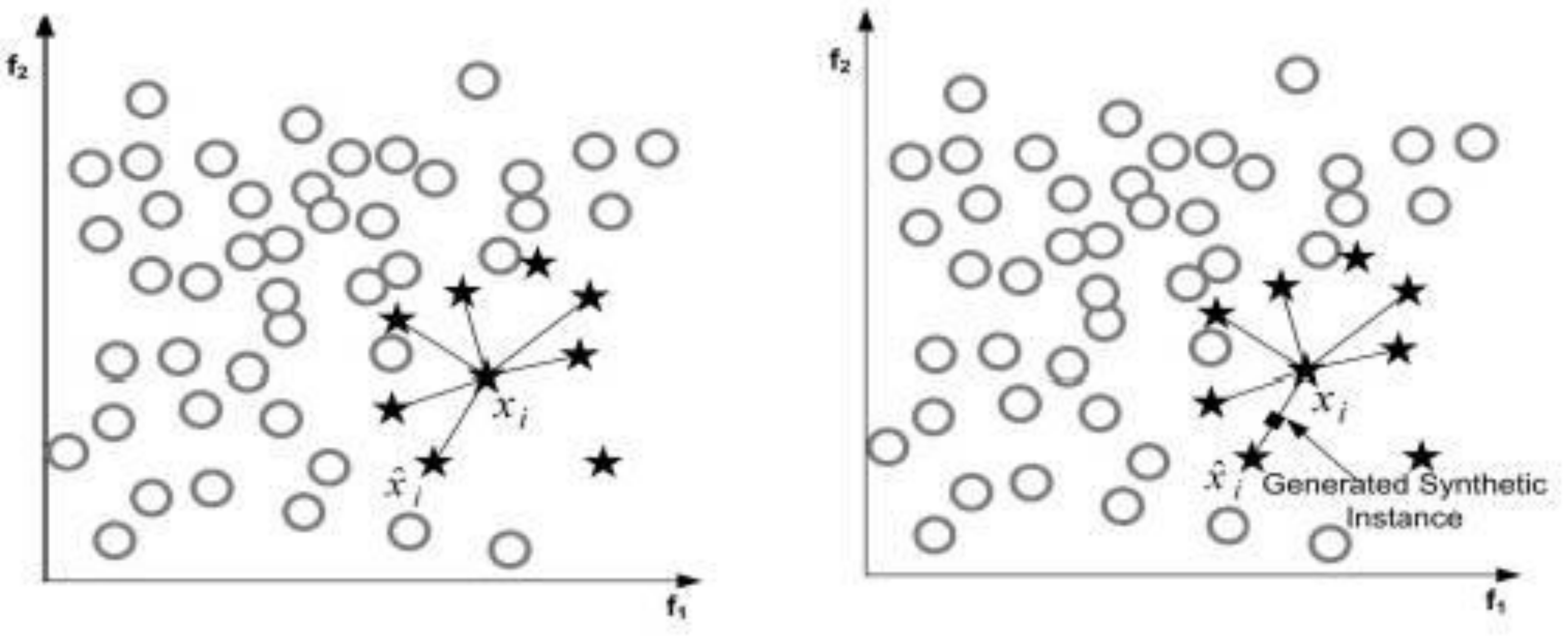
Introduction

Class imbalance problem

- data-resampling
  (oversampling、undersampling...)
- algorithm-level methods
  (the cost-sensitive learning algorithms)

SMOTE算法

(Synthetic Minority Oversampling Technique)



(a) (b)

★ randomly select a minority point and find its kNN point in the minority. Then randomly mark a point from kNN.

★ Randomly select a point on the line between a and b as the newly synthesized minority sample

Introduction

Contributions:

A novel cost-sensitive active learning framework that combines uncertainty and diversity measures for sample selection is proposed for imbalanced learning.

A novel algorithm that measures the diversity of examples is proposed, where the K-means clustering is firstly used to scatter examples.

A set of experiments are conducted on several class-imbalanced datasets to confirm the advantages of the proposed method over some state-of-the-art methods.

Method

1、 Uncertainty Measure

the labeled dataset is $\mathcal{L} = \{x_1, x_2, . . ., x_n\}$

the unlabeled dataset is $\mathcal{U} = \{x_{n+1}, x_{n+2}, \dots, x_{n+m}\}$

$$y_i = \begin{cases} -1, & \mathcal{F}(x_i) < \theta \\ 1, & \underline{\mathcal{F}(x_i)} \geq \theta \end{cases}$$

$$unc_i = |\mathcal{F}(x_i) - \theta|$$

(the model function)

Smaller distance means that the prediction result on is more uncertain.

Method

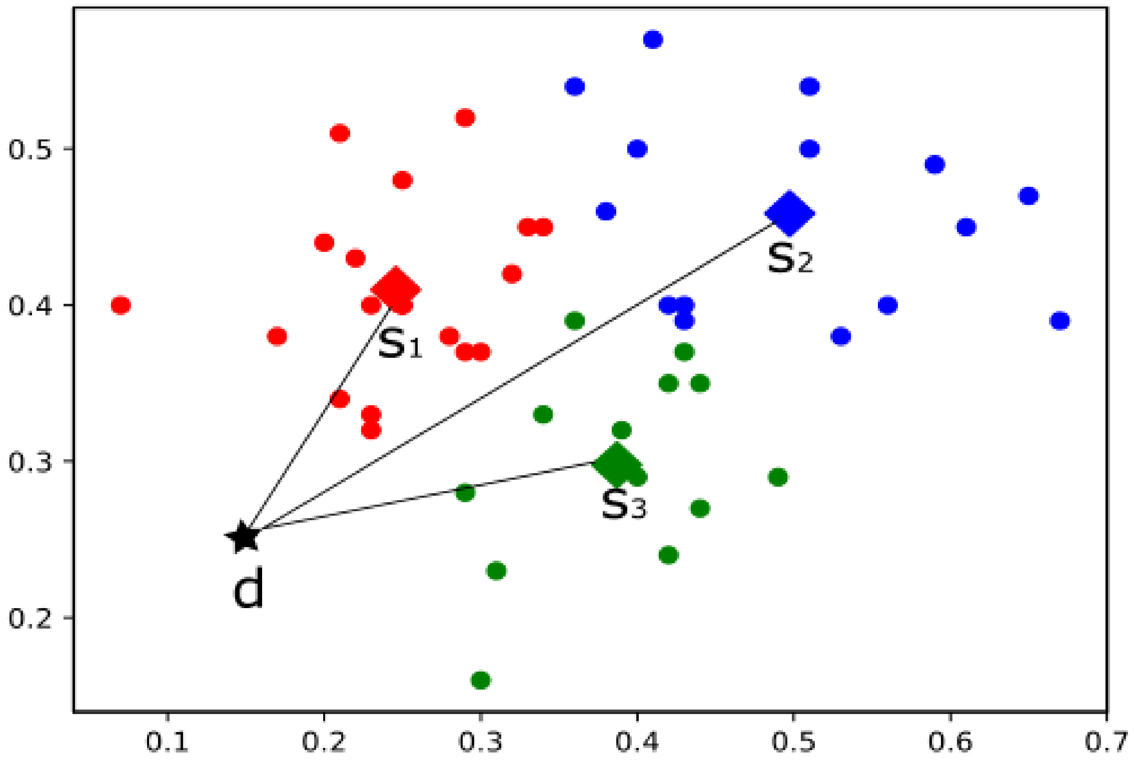
2、 Diversity Measure

(1) $\mathcal{D} = \{d_1, d_2, \ldots, d_t\} \subseteq \mathcal{U}$

$\mathcal{L} = \{l_1, l_2, \ldots, l_l\} \xrightarrow{\text{k-means}}$ k center points (Cj)

(2) $$D_d = \sum_{j=1}^{k} \left\| c_j - \mathrm{d} \right\|^2$$

(3) choose samples in whose distance is far from all center points to make our selected samples diverse.



The distance D_d is the sum of the distance of s_1d, s_2d, and s_3d.

Algorithm Imbalanced Active Learning (IAL)

Input: labeled dataset $\mathcal{L}$, Unlabeled data pool $\mathcal{U}$, parameters $m_s$and m_l

Output: Classifier $\mathcal{F}$

1. Initialize k of k-means cluster algorithm, t,n

2. REPEAT // In each iteration j

2.1 Calculate the class weight $\omega_v = \frac{m_l}{m}$ and $\omega_\mu = \frac{m_s}{m}$.

Classifier $\mathcal{F}_{\mathrm{j}} \leftarrow$Train $(\omega_v, \omega_\mu, \mathcal{L})$

Obtain the center points $\{c_i\}$ of $\mathcal{L}$ using k-means cluster algorithm

2.2 Calculate the uncertainty of instances by Eq. (2)

Get the most t uncertain samples ($\mathcal{D}$)

FOR d in $\mathcal{D}$ **DO**

Calculate the diversity of instances by Eq. (3)

Get the most n diverse samples in $\mathcal{D}$.

2.3 Query the labels of the n samples, add the labeled n samples to the labeled dataset $\mathcal{L}$, update $\mathcal{U}$.

UNTIL the performance of $\mathcal{F}$ is satisfied.

3. RETURN $\mathcal{F}$

$$unc_i = |\mathcal{F}(x_i) - \theta|$$

$$D_d = \sum_{j=1}^{\mathrm{k}} \|c_j - \mathrm{d}\|^2$$

Experiments: Dataset

Table 1. Characteristics of the six datasets

dataset	size	#attribs	#major/#minor	ratio
Yeast1	1484	8	1055/429	2.46
Vehicle2	846	18	628/218	2.88
Ecoli1	336	7	259/77	3.36
Segment0	2308	19	1979/329	6.02
Page-block0	5472	10	4913/559	8.79
Abalone9-18	731	8	689/42	16.4

Experiments: Evaluation Metrics

Table 2. Confusion matrix

	Predict Positive	**Predict Negative**
True Positive	TP	FN
True Negative	FP	TN

Specificity:

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

Precision:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

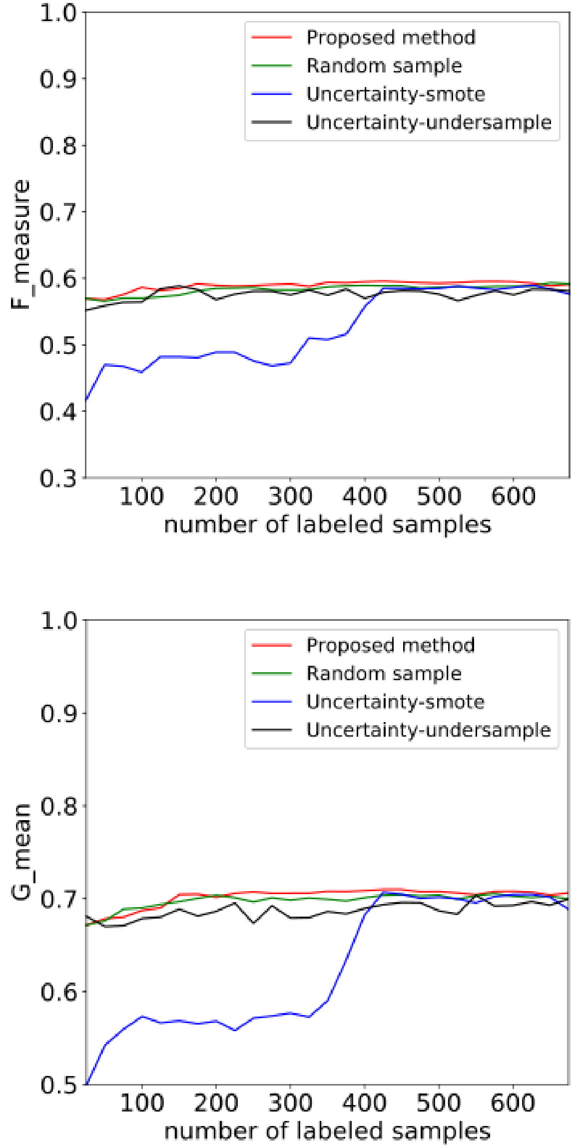
F_measure: suggested in [18], which mixes *recall* and *precision* as an average:

$$\boxed{F\_measure} = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (7)$$

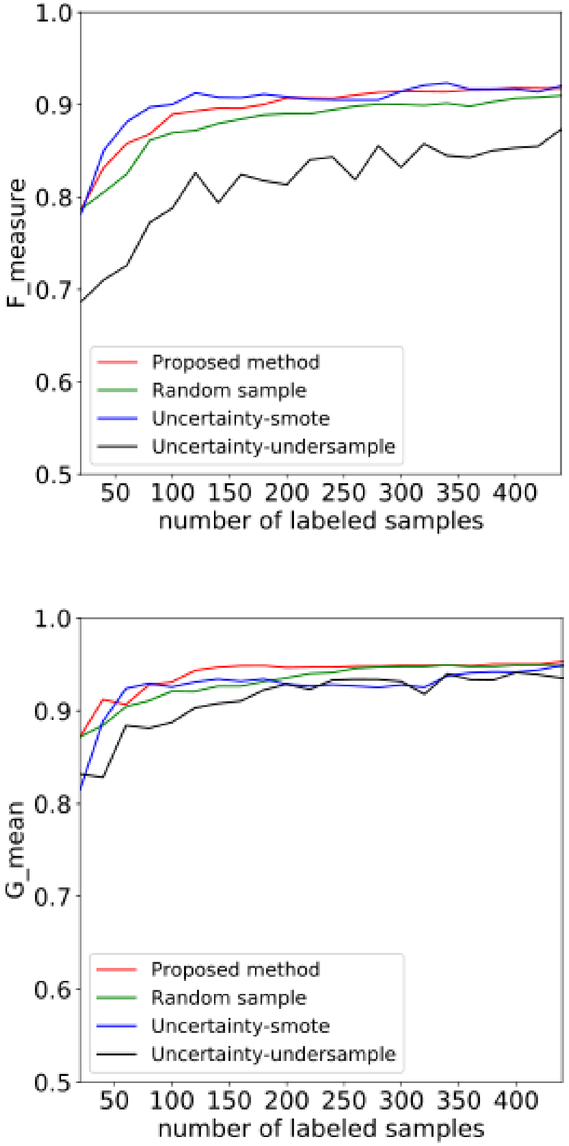
G_mean: When the performance of both classes is concerned, both *specificity* and *recall* are expected to be high in the meantime. It defined as :

$$\boxed{G\_mean} = \sqrt{Specificity \times Recall} \quad (8)$$

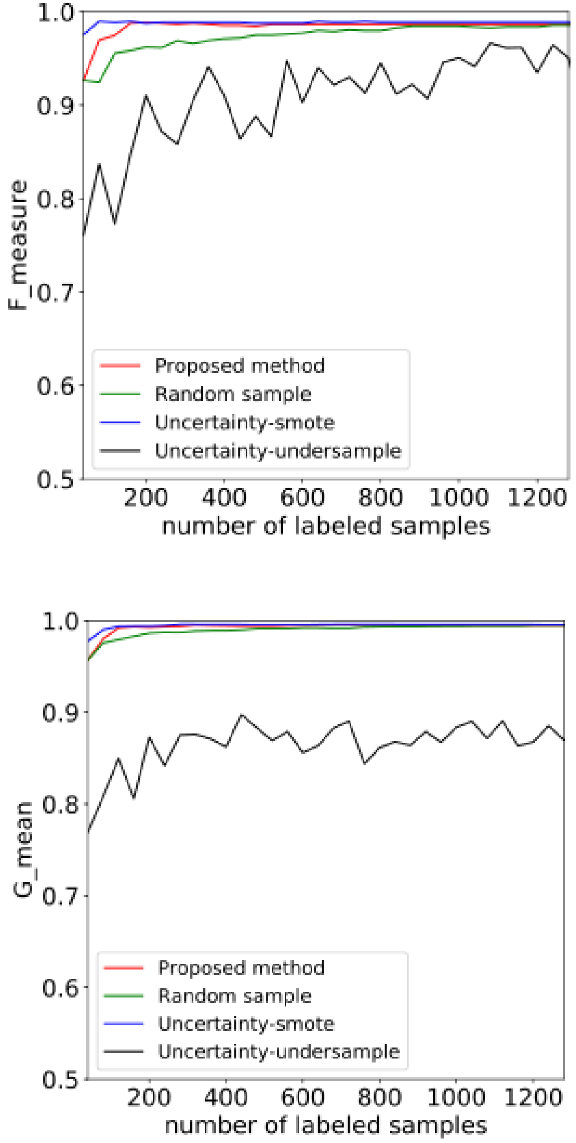
Experiments: Results



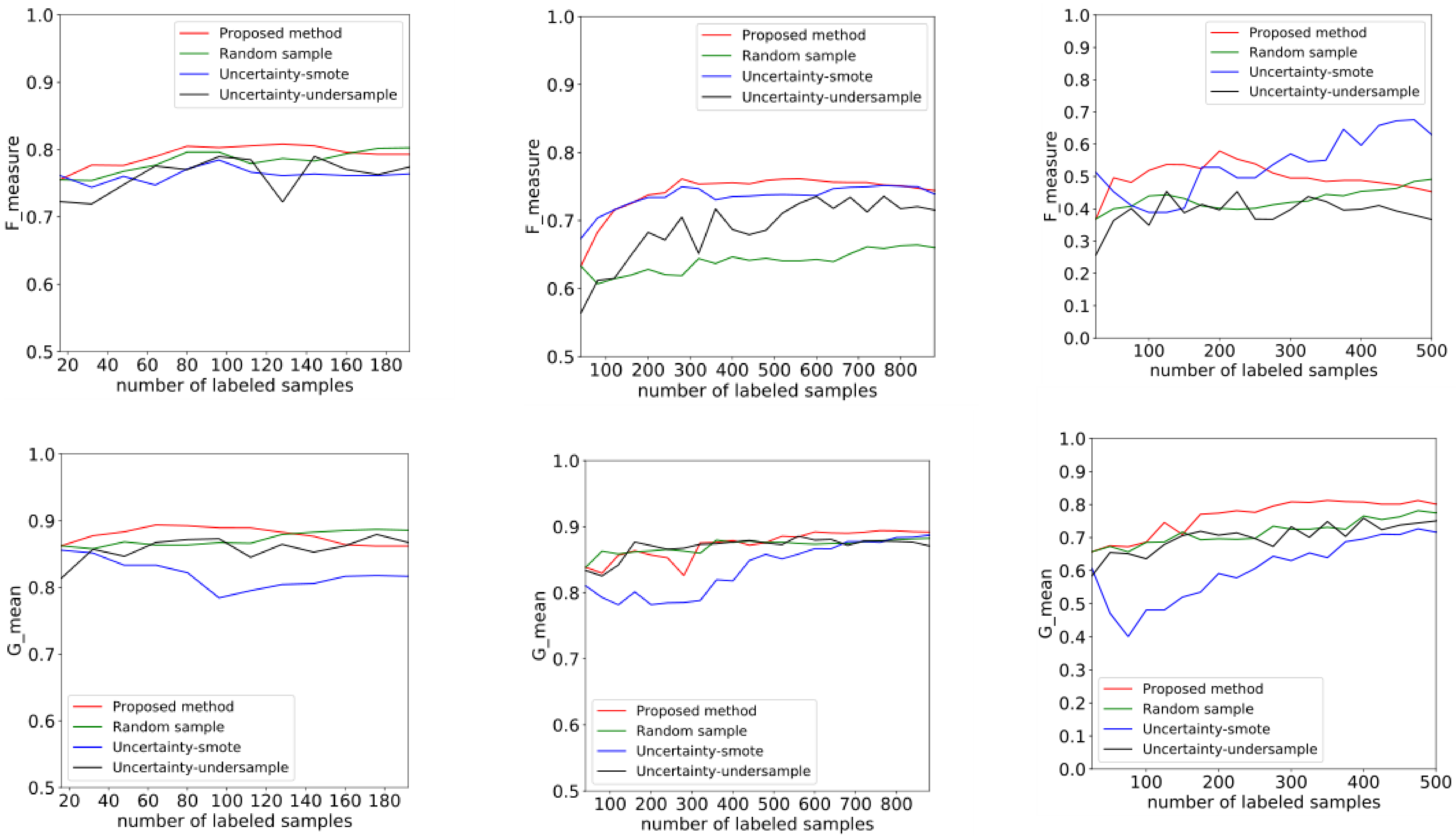
dataset Yeast1



Vehicle2



Segment0



Ecoli1 Page-block0 Abalone9-18

Thanks