Sensitivity analysis of initial classifier accuracy in fuzziness based semi-supervised learning

Muhammed Jamshed Alam Patwary

PhD Research Fellow, Big Data Institute

College of Computer Science and Software Engineering

Shenzhen University

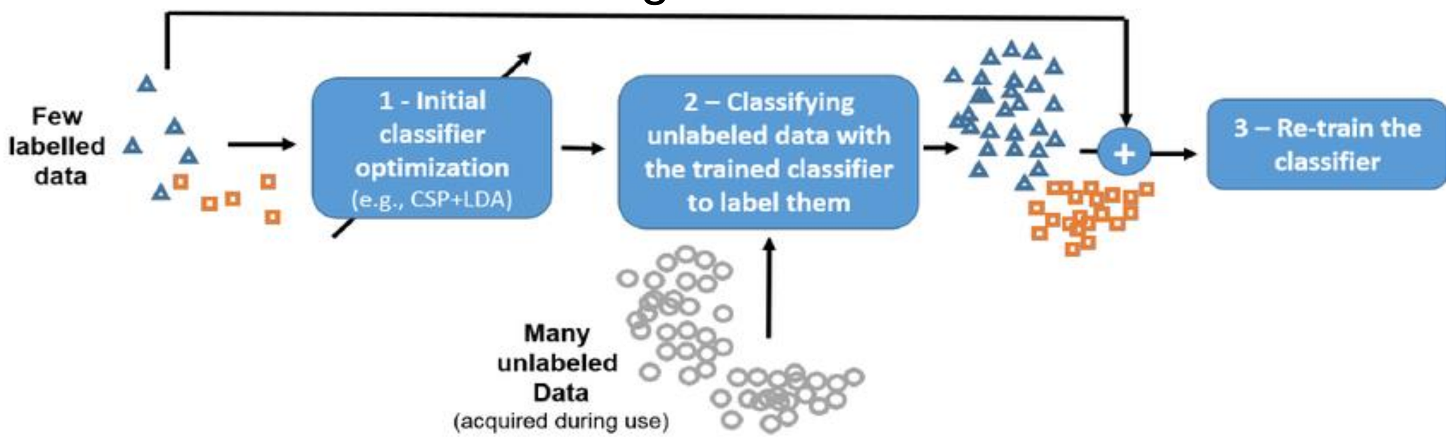
Outlines

- Motivation
- Main contribution
- Proposed algorithm
  - Main idea
  - The pseudo-code
- Experimental results and analysis
  - Experimental results of 1^{st} experiment
  - Experimental results of 2^{nd} experiment
- Conclusions and Future works

Motivation

The Traditional View:

- Labeled instances are difficult to get
  - Expensive and time consuming to obtain.
  - They require the effort of experienced human annotator.
- Unlabeled data is cheap
- **Semi-supervised learning** is a class of supervised learning tasks and techniques that also make use of unlabeled data for training



Motivation

- Why Semi-supervised learning?
- The learning problem
  - Goal: Using both labeled and unlabeled data to build better learners, then using each one alone.
- **Fuzziness** refers to the inexactness existing in an unclear event.
- **The goal** is to analyze the sensitivity on initial classifier accuracy in fuzziness based semi-supervised learning.

Main contribution

1. It is experimentally shown that if we add the low fuzziness instances from test dataset to the original training dataset then its testing accuracy can be improved.
2. The phenomenon pointed out in (1) is theoretically explained based on the theory of learning from noisy data.
3. It is found that, regarding the phenomenon in (1) and (2), the initial accuracy of the base classifier has a significant impact on the improvement in accuracy.
4. It is experimentally observed that the maximum improvement of accuracy of the classifier is attained when the initial accuracy is approximately between 70%-80%.

Proposed Algorithm-main idea

In semi-supervised setting, when initial classifier is used to classify huge amount of unlabeled data several events may turn out.

- When initial classifier's accuracy is very low, for instance about 50%, if we use this classifier to predict unlabeled data, the predicted labels may have some noises.
- When initial classifier's accuracy is medium, for example around 75%, if we use this classifier to predict unlabeled data, then it works very well, the predicted labels may have very few noises.
- When initial classifier's accuracy is very high, for example around 95%, if we use this classifier to predict unlabeled data, then it may generate a few wrongly-predicted labels.

Algorithm 1 Fuzziness based semi-supervised learning algorithm

Input: Dataset, # hidden layer Z, # node in each layer Q.

Output: Maximum improvement of accuracy over the initial accuracy, corresponding initial accuracy.

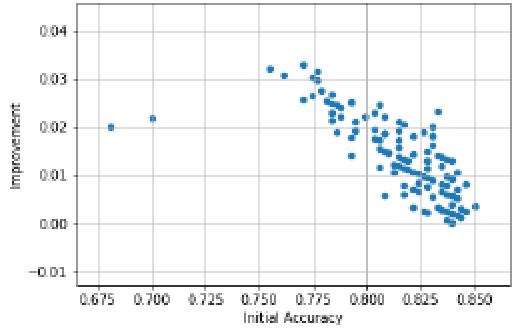
```
Randomly partition the dataset into a training dataset X_tr and a testing dataset X_te.
x ← 1
k ← 1
while x <= Z do
    n ← 1
    while n <= Q do
        Train the classifier C according to a training algorithm.
        Get the training accuracy tr_accB
        Get the fuzzy vector A_i = {θ_1, θ_2, ··· , θ_n} for each sample in testing set by classifier C.
        Calculate the fuzziness P(A_i) of each sample in testing set by: P(A_i) = −(1/n) Σ_{i=1}^{n} θ_i logθ_i + (1 − θ_i) log(1 − θ_i)
        Sort the samples by the fuzziness P(A_i), and group testing set X_te into three fractions: X_te low, X_te medium and X_te high.
        Get new training set X_tr new by adding the low-fuzziness samples X_te low to the original training set X_tr.
        Retrain a new classifier C_new according to the given training algorithm with X_tr new.
        Again record the training accuracy tr_accA by classifier C_new with X_tr new
        Record diff[k] = tr_accA − tr_accB
        n = n + 1
        k = k + 1
    end while
    x = x + 1
end while
Find the maximum of diff and the corresponding initial accuracy
```

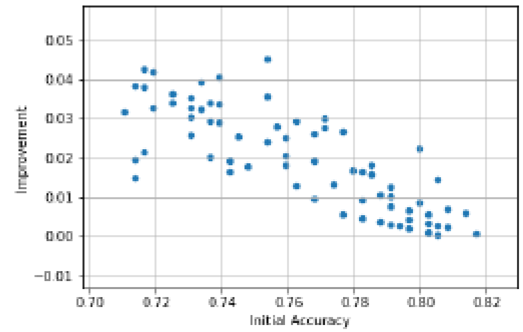
Experimental results and analysis

Table 1: Dataset description

Dataset	# Instances	# Features	# Classes
Blood Transfusion Service Center Dataset	749	4	3
Indian Liver Patient Dataset (ILPD)	582	10	2
Phishing Dataset	1354	9	3
Pima-indians-diabetes	769	8	2
HIGGS-30000 dataset	29841	28	2
letter-recognition	20000	16	26
magic04 dataset	19020	10	2
waveform	5000	21	3
vehicle	846	18	4
Ecoli	336	7	8
Sonar	208	60	2
Parkinson	195	22	2
YALE dataset	165	1024	15
ORL dataset	400	1024	40

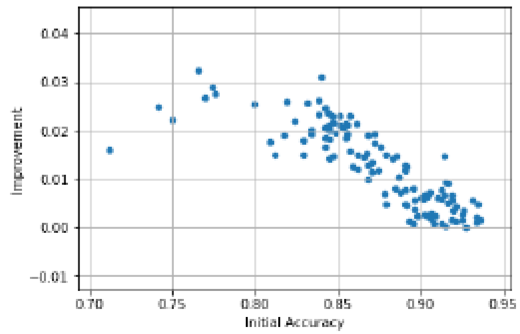
Experimental results of 1st experiment when ELM is used as initial classifier.

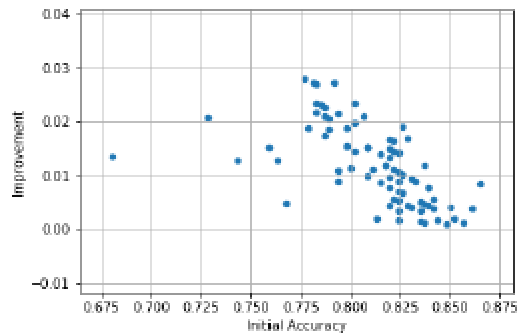




(a) Blood Transfusion Service Center Dataset

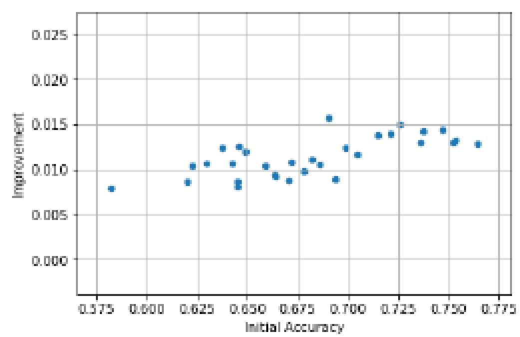
(b) Indian Liver Patient Dataset

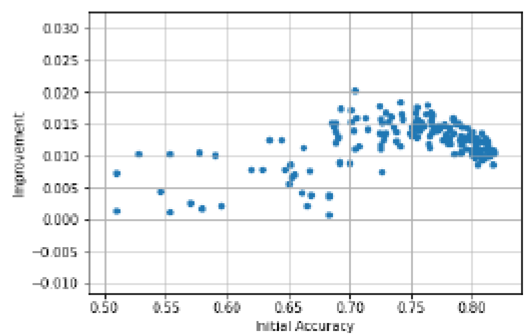




(c) Phishing Dataset

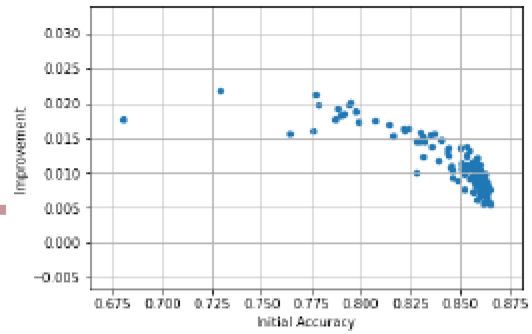
(d) Pima-indians-diabetes

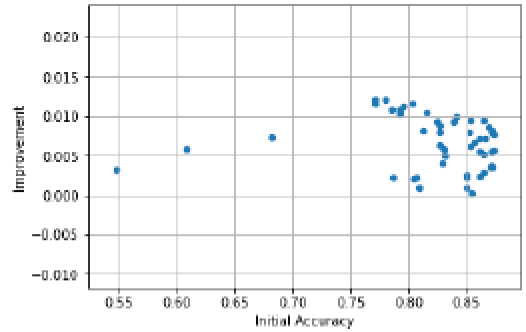




(e) HIGGS-30000 dataset

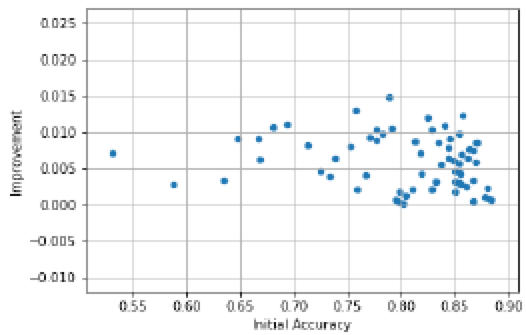
(f) Letter recognition

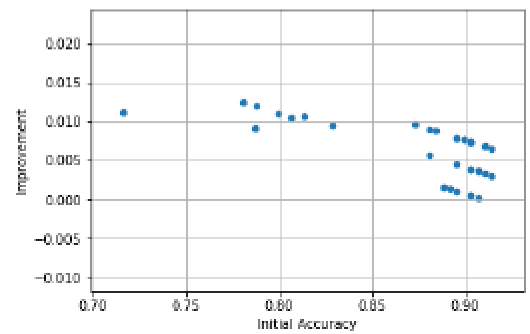




(g) magic04 dataset

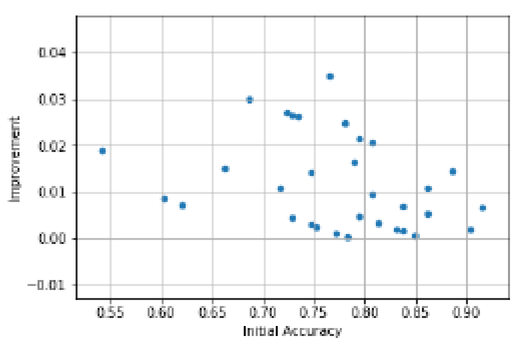
(h) waveform dataset

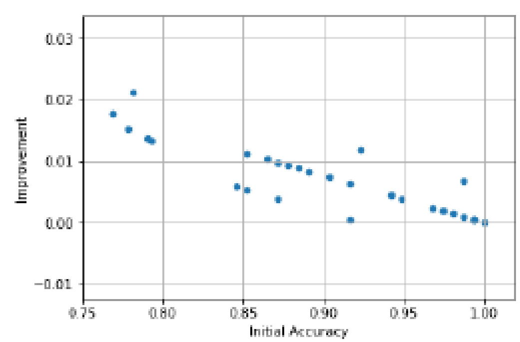




(i) Vehicle dataset

(j) Ecoli

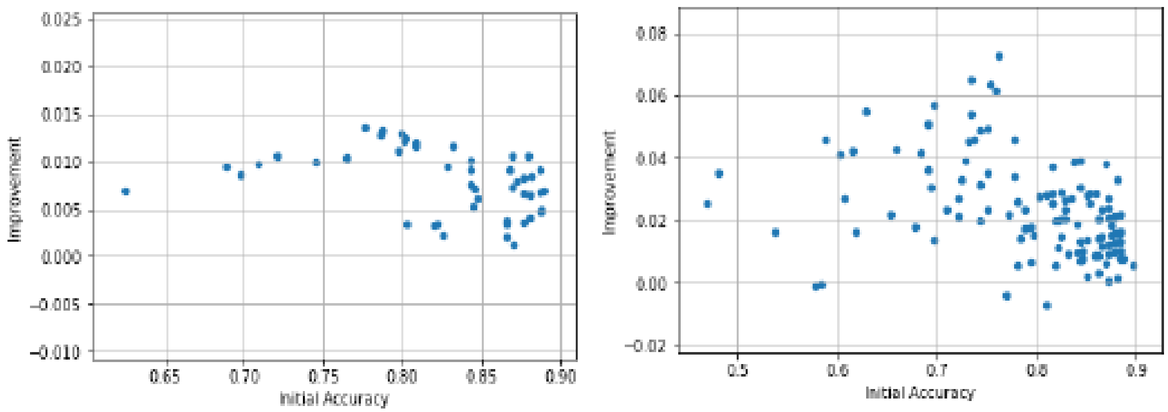




(k) Sonar dataset

(l) Parkinson dataset

Experimental results of 1st experiment when ELM is used as initial classifier

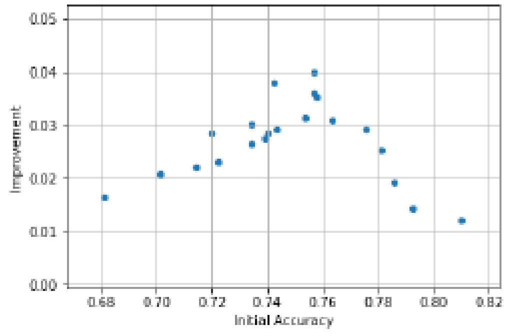


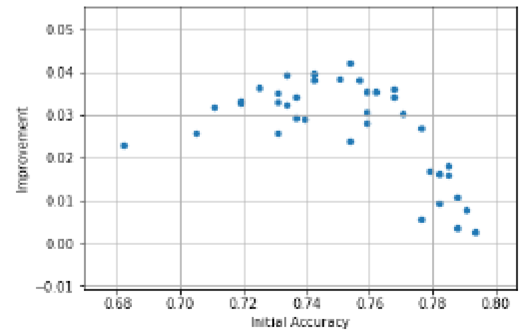
(m) YALE dataset

(n) ORL dataset

Figure 4: Results of 1st experimental setup (ELM used as base classifier) (cont.)

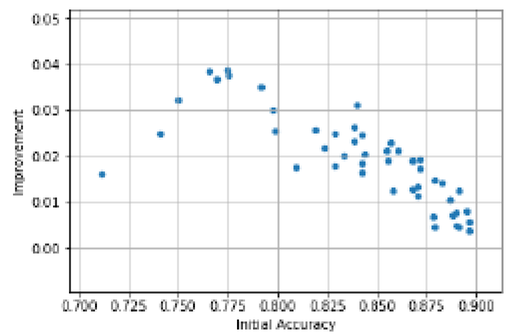
Experimental results of 2nd experiment when NN is used as initial classifier.

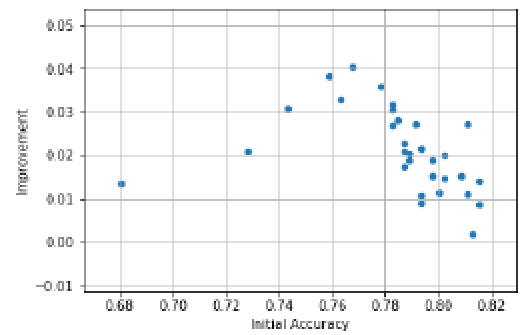




(a) Blood Transfusion Service Center Dataset

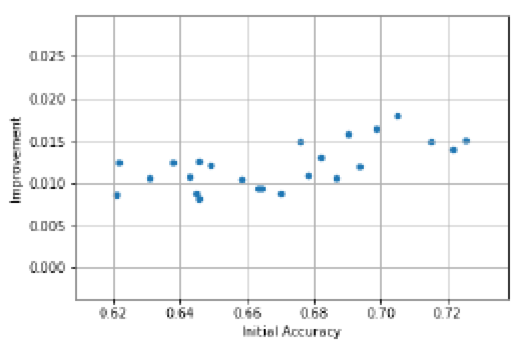
(b) Indian Liver Patient Dataset

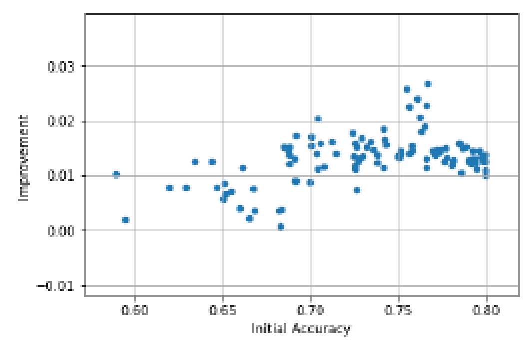




(c) Phishing Dataset

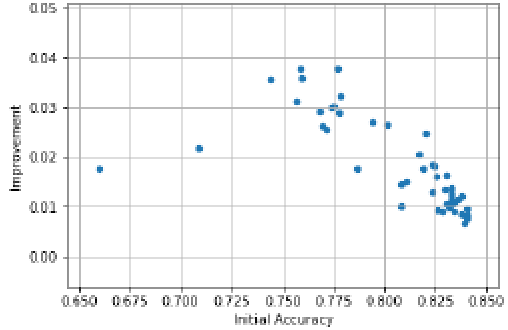
(d) Pima-indians-diabetes

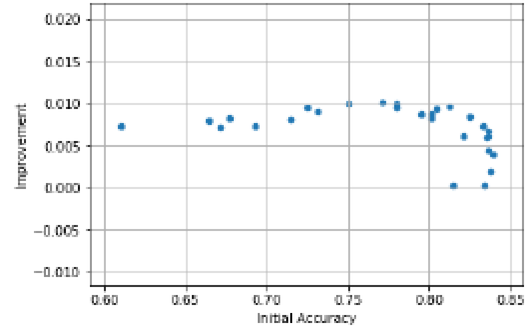




(e) HIGGS-30000 dataset

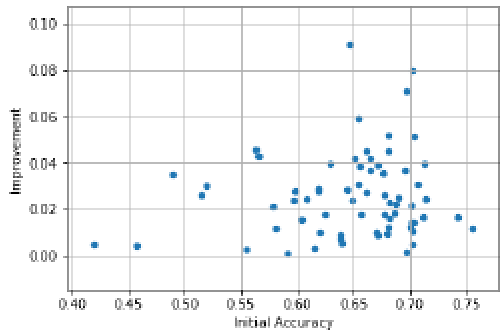
(f) Letter recognition

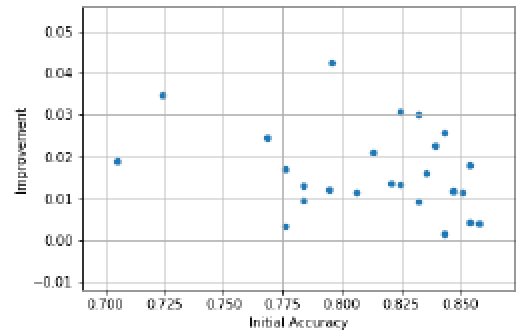




(g) magic04 dataset

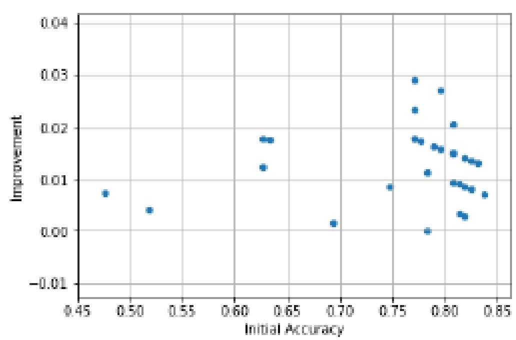
(h) waveform dataset

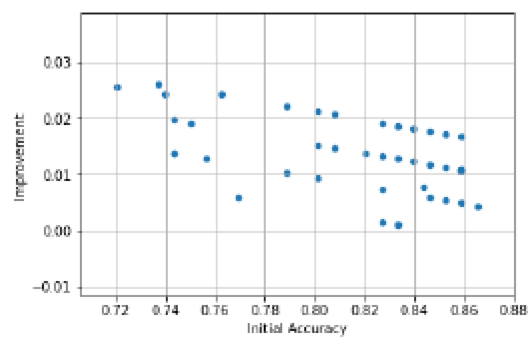




(i) Vehicle dataset

(j) Ecoli

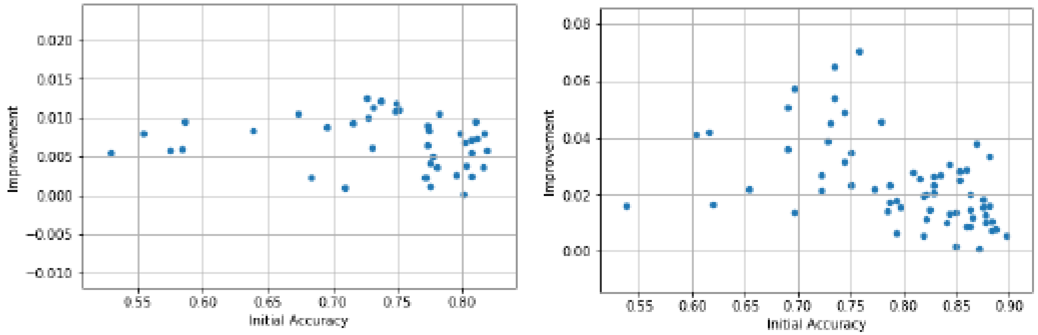




(k) Sonar dataset

(l) Parkinson dataset

Experimental results of 2nd experiment when NN is used as initial classifier



(m) YALE dataset

(n) ORL dataset

Figure 3: Results of 2nd experimental setup (NN used as base classifier) (cont.)

Experimental Results

Table 2: ELM used as initial classifier

Dataset	Accuracy before adding low fuzzy samples	Accuracy after adding low fuzzy samples	Improvement of accuracy
Blood Transfusion Service Center Dataset	0.770089286	0.80291971	0.03283
Indian Liver Patient Dataset	0.753581662	0.79859485	0.045013
Phishing Dataset	0.765721332	0.79818365	0.032462
pima-indians-diabetes	0.776521739	0.80451957	0.027998
HIGGS-30000	0.690013405	0.70577686	0.015763
letter-recognition	0.703923077	0.72418966	0.020267
magic04 dataset	0.729111057	0.75085722	0.021746
waveform	0.78	0.7921628	0.012163
vehicle	0.788461538	0.80327869	0.014817
ecoli	0.780597015	0.79310345	0.012506
sonar	0.765060241	0.8	0.03494
parkinson	0.781282051	0.80248521	0.021203
yale	0.776	0.7896628	0.013663
ORL	0.7625	0.83540462	0.072905

Experimental Results

Table 3: NN used as initial classifier

Dataset	Accuracy before adding low fuzzy samples	Accuracy after adding low fuzzy samples	Improvement of accuracy
Blood Transfusion Service Center Dataset	0.756785714	0.796569343	0.039784
Indian Liver Patient Dataset	0.753581662	0.795594848	0.042013
Phishing Dataset	0.774352651	0.813229062	0.038876
pima-indians-diabetes	0.767391304	0.80789819	0.040507
HIGGS-30000	0.704814567	0.72273553	0.017921
letter-recognition	0.766769231	0.793538773	0.02677
magic04 dataset	0.758694492	0.796518722	0.037824
waveform	0.77129	0.781452797	0.010163
vehicle	0.646449704	0.737704918	0.091255
ecoli	0.795970149	0.838275862	0.042306
sonar	0.771084337	0.8	0.028916
parkinson	0.737179487	0.763313609	0.026134
yale	0.726	0.738562797	0.012563
ORL	0.7575	0.828014624	0.070515

Conclusions and Future works

- In this study, a new aspect of semi-supervised learning technique was explored to improve the performance of a classifier by using divide-and-conquer strategy.
- One of our future works is to establish a robust mathematical model to explain why low-fuzziness samples have the enhanced impact on the learning performance.
- We will study the impacts of selecting different initial classifiers on the learning performance of SSL.
- We will conduct a detailed survey on SSL techniques.