Centrality in Active learning for Graph based semi-supervised learning

Ankit Shrivastava, Krishna Manglani
Computational and Data Sciences
Indian Institute of Science, Bangalore, India
ankitvaibhava@gmail.com, krish2100@gmail.com

***Abstract*—Today, large amount of unlabelled data is available and preparing them into a labelled data is extremely expensive. Several Semi-Supervised Learning (SSL) have been proposed which learns from labelled data as well as from unlabelled data. For SSL method to predict data with high accuracy we need labelled portion of data with most informative instance. Active learning can be used to selectively label instances by interactively selecting most informative instances based on certain instance-selection criteria. In this paper, we have studied the various Active learning instance-selection methods and their effects on Graph based SSL algorithm to predict the labels for remaining unlabelled data. We have also explored the role of centrality in Active learning with GSSL as a model in query stage. We have also showed that clique-overlap as one of the robust and good metric for centrality in instance selection in Active learning.**

I. Introduction

Training a model from available labelled set of data is one of the most popular paradigm of machine learning.The labelled data are usually prepared using human inputs. However in most of the case we have only unlabelled set of data and preparing labelled sets from these unlabelled sets using human inputs is expensive. For this semi-supervised based learning has been proposed which use a set D which contains both labelled instances and unlabelled instances i.e. $\{D_l\ D_u\}$ for training. Graph based SSL (GSSL) are sub-class of SSL where each data sample $x_i \in D$ is represented by a vertex in a weighted graph with the weights providing a measure of similarity between vertices. The seed labels are injected at to the nodes representing label instances and probable labels of unlabelled nodes of graph are predicted by using label propagation-like methods.

We will have to prepare set of labelled and unlabelled instances before we execute any GSSL algorithm. However, major challenge is to select optimum number of most informative instances that can be labelled interactively, which are then used to label other unlabelled instances using any GSSL method.

Active learning aims to minimize the cost of labelling by selecting most informative instance and then labeling them. The informativeness of any instance can be calculated from utility metrics. The utility metrics calculation can be classified into two parts:

- Utility metric merely based on uncertainty of IID.
- Utility metrics further taking into account instance correlations.

In first case i.e metric based on uncertainty of IID selection criteria only depend on the uncertainty values computed with respect to each individual instances own information.Using this metric we are likely to select a candidate set only containing the most uncertain instances from an individual samples perspective. However in this case these instances may contain redundant knowledge and therefore do not form an ideal candidate set. To take the sample redundancy into consideration we will have to consider uncertainty metrics based on instance correlation. It utilizes some similarity measures to discriminate differences between instances. Hence we select the instances by calculating utility metrics by balancing the uncertainty and diversity.

There are many proposed methods available to calculate the uncertainty and diversity. We have considered few of them and tried all possible combinations to find out which combination can lead to better results. Hence we first select set of instances using our utility metrics. Interactively we label these selected instances and then execute GSSL code to label rest of the unlabeled instances.Then error is evaluated with actual labels of the unlabeled data.Hence we are trying to compare how active learning can improve prediction of labels.

While exploring diversity in Active Learning we have introduced Clique Overlap Centrality and use it to compare the results with other centrality. Cross-clique centrality of a single node, in a complex graph determines the connectivity of a node to different cliques. A node with high cross-clique connectivity facilitates the propagation of information or disease in a graph. Cliques are subgraphs in which every node is connected to every other node in the clique. The cross-clique connectivity of a node v for a given graph $G := (V, E)$ with $|V|$ vertices and $|E|$ edges, is defined as X(v) where X(v) is the number of maximal cliques to which vertex v belongs. This measure was used in [1] but was first proposed by Everett and Borgatti in 1998 where they called it clique-overlap centrality. Hence in short it is equal to $\Sigma_k(is.in.maximal.clique(v, k))$.

The rest of the report is organised as follows: In Section-2, Related Work of our hybrid execution is summarised. In

Section-3, we talk about our hybrid implementation methodology wherein we describe all the steps in details. In Section-4, experiment setup and results are presented along with analysis. We finally conclude in Section-5.

II. Related Work

The area of active learning is not new and good documentation is available in references [2] and [3]. However, using centrality in active learning to improve the accuracy of classification was recently proposed in [4]. The author considers betweenness and closeness centrality for generating the seed data which would be used in Label Propagation algorithm for predicting labels of unlabelled data.

III. Methodology

In this section, we provide a complete description of the algorithm that we have employed in Active learning. Following steps are executed for algorithms:

- Initialize $\{D_L\ D_U\}$ randomly.
- Using $\{D_L\ D_U\}$ find $P_\theta(\hat{y} \mid x)$
- Using $P_\theta(\hat{y} \mid x)$ find uncertainty
- Now calculate diversity by exploiting graph structure
- Calculate utility metrics using diversity and uncertainty.
- Choose k instances with top utilities and then interactively label them
- Finally add those instances to D_L and remove it from D_U.

Algorithm 1. by Yifan Fu [2] explains general scheme of active learning. To compute uncertainty we have used following schemes:

- Least-Confidence [5]: $x^*_{LC} = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y} \mid x)$
- Margin [6]: $x^*_M = \underset{x}{\operatorname{argmin}}\ P_\theta(\hat{y}_1 \mid x) - P_\theta(\hat{y}_2 \mid x)$
- Entropy: $x^*_E = -\Sigma_i P_\theta(\hat{y}_i \mid x)\, log P_\theta(\hat{y}_i \mid x)$

Now we estimate Cross-clique centrality. Algorithm 2. explains the general scheme of Cross-clique centrality proposed by Mohammad Reza Faghani [1] . We also compare the results with other following Centrality schemes:

- Degree Centrality : $C_D(v) = deg(v)$
  It represents number of links incident upon a node.
- Closeness Centrality: $H(x) = \Sigma \frac{1}{d(y,x)}$
  It represents length of their shortest paths.
- Betweenness Centrality: $C_D(v) = deg(v)$
  It represents number of times a node acts as a bridge along the shortest path between two other nodes
- Page-rank Centrality: $x_i = \alpha \Sigma_j a_{ji} \frac{x_j}{L(j)} + \frac{1-\alpha}{L(N)}$
  It represents the quality of vertex and edges. It is the probability of random walker to reach at the given vertex after walking infinite steps. It is a special case of eigen vector centrality.

Algorithm 1 General Process of Active learning

```
Initialise labeled instance set D_L and Unlabelled Instance set D_U andd size of required training set m
while training size ≤ m do
    θ ← learn a model based on D_L
    D_U ← D \ D_L
    for each instance x_i in D_U do
        u_i ← u(x_i, θ)
    end for
    x* ← argmax_i (u_i)
    D_L ← D_L ∪ x*
    D_U ← D_U \ x*
    θ ← update the model based on D_L
end while
```

Algorithm 2 Finding well-cross-connected nodes

```
Input: a set of cliques C = {c_1, c_2, ....c_k}; a set of counters Q = {u_1, u_2, ....u_k}
Output: Ordered set of well-crossed-connected nodes stored in List;
for i = 1 → k do
    for each user i in c_i do
        u_j = u_j + 1
    end for
end for
List ← sort Q in non increasing order
return List
```

Now with these new $\{D_L\ D_U\}$ we execute GSSL algorithm and try to predict labels for D_U. We have used three different GSSL algorithms

- Label Propagation [7]
- Local and Global Consistency [2]
- Modified Adsorption [8]

IV. Experiments and Results

A. Experiment Setup

Experiments are done against following sets of independent variables:

- Labelling budget is increased in steps of 15 till it reaches 30 % of total data set from zero.
- Uncertainty type = $\{leastconfidence\ \ , margin\ \ , entropy\}$
- Centrality type = $\{betweenness\ \ degree\ \ closeness\ \ pagerank\ \ clique\_overlap\}$
- GSSL = $\{LP\ \ MAD\ \ LGC\}$

Data set are chosen from Analysis of benchmark chapter 21 which are available at olivier.chapelle.[1]. Following data sets are chosen $\{g241c\ \ g241n\ \ Digitl\ \ USPS\ \ COIL\ \ COIL_2\ \ BCI\}$

The datasets g241c, g241n follow cluster assumption but do not follow manifold assumption. The dataset digit1 follows

[1] http://olivier.chapelle.cc/ssl-book/benchmarks.html

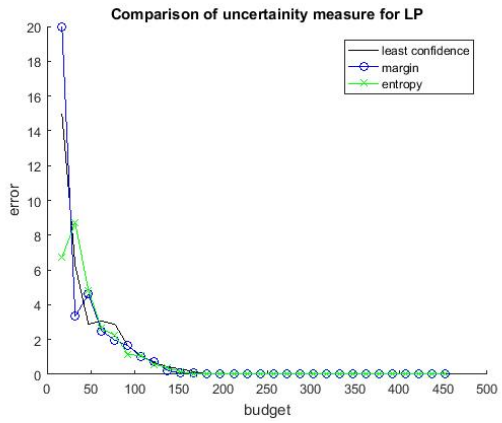

Fig. 1. Error v/s Budget for data set Digit1

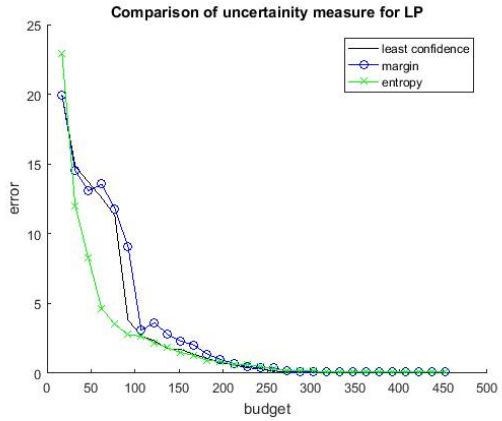

Fig. 2. Error v/s Budget for data set USPS

manifold assumption but not cluster assumption whereas the dataset USPS holds both manifold as well as cluster assumption.

The effect of uncertainty and centrality was observed on accuracy of active learning model. The true test error was used as the measure of evaluation as the labels are already known in the training data and GSSL is a transductive scheme.

Finally, the performance of proposed Active learning model (least confidence with clique overlap centrality) was tested on digit1 dataset for different GSSL methods.

B. Results

Figures (1-7) shows comparisons of error v/s budget when either of uncertainty (Least-confidence,margin,Entropy) is used for different data sets.

Figure 8 shows the comparison of errors for different data sets and for different uncertainty measures for given budget = 10.

Figure 9 shows the comparison of errors for different data sets and for different uncertainty measures for given budget = 100.

As we can see from the figures 8 and 9 , the active learning helps to provide better accuracy estimates for the datasets

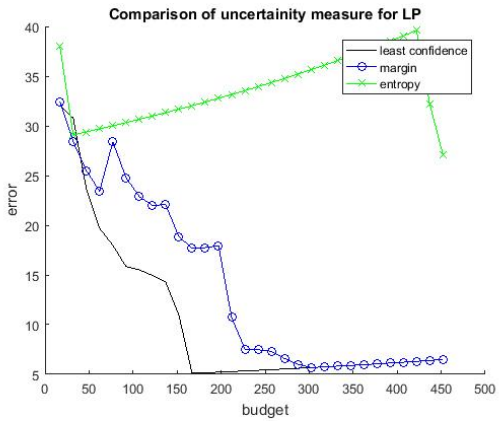

Fig. 3. Error v/s Budget for data set COIL2

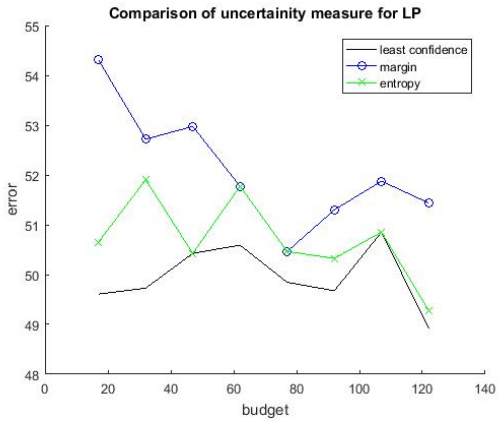

Fig. 4. Error v/s Budget for data set BCI

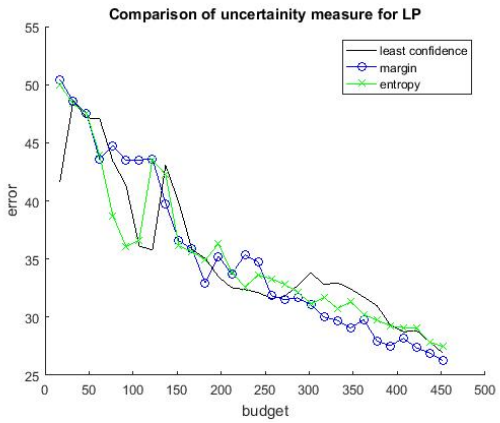

Fig. 5. Error v/s Budget for data set g241c

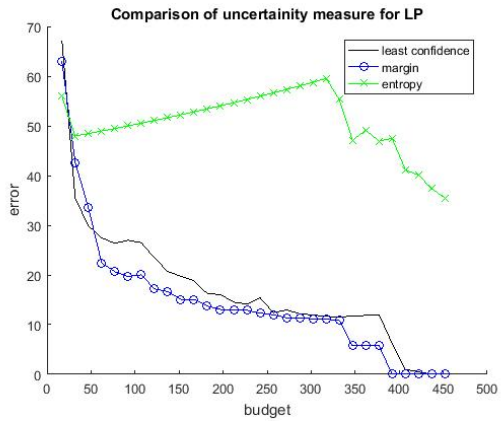

Fig. 6. Error v/s Budget for data set COIL

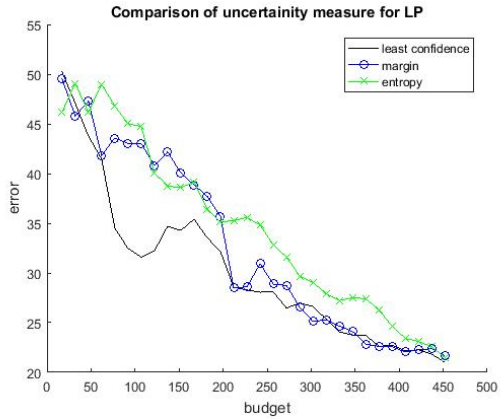

Fig. 7. Error v/s Budget for data set g241n

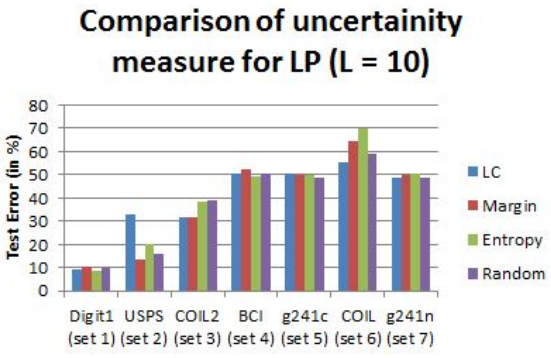

Fig. 8. Error v/s Budget for data set BCL

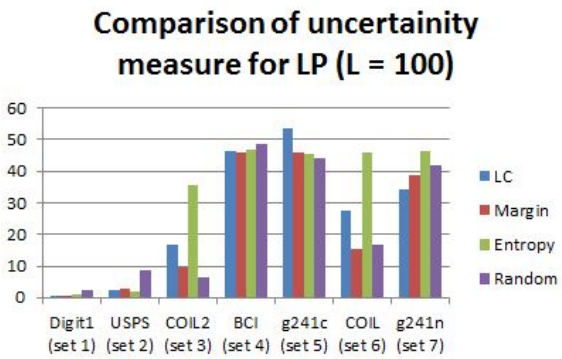

Fig. 9. Error v/s Budget for data set BCL

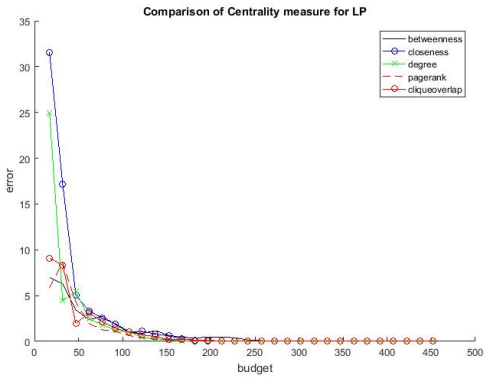

Fig. 10. Error v/s Budget for data set Digit1

which follow manifold assumption like USPS,Digit1 datasets since label propagation works well for manifold data. For others, we get the accuracy of the same order as random, this again could be argued from the same reason that they do not have manifold structure and label propagation do not work well on them.

Figure (10-16) shows comparison of error v/s budget when different centrality measures (Betweenness , Closeness ,Degree Page-rank, Clique-overlap) are used to calculate diversity for different data sets.

Figure 17 shows the comparison of errors for different data sets and for different centrality measures for given budget = 10.

Figure 18 shows the comparison of errors for different data sets and for different centrality measures for given budget = 100.

The figures 17 and 18 shows that the proposed clique overlap centrality method turns out to perform better than other centrality measures used in making utility function.

It was also seen that entropy model is unstable with different data distributions. Hence we choose least confidence as the uncertainity representation for our active learning model.

The figure 19 shows the performance of the proposed active learning model for different GSSL methods.

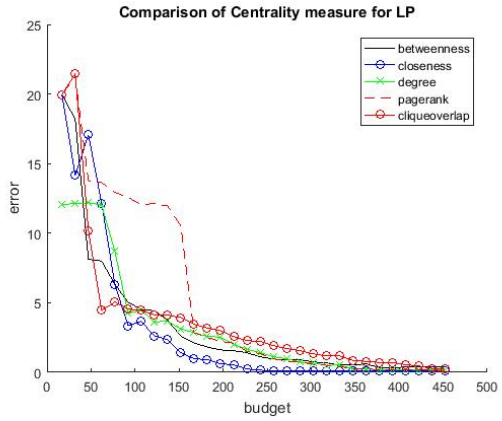

Fig. 11. Error v/s Budget for data set USPS

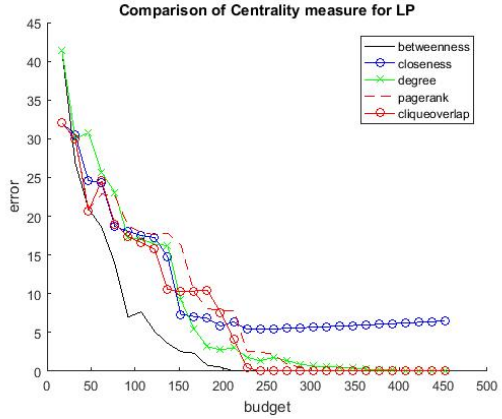

Fig. 12. Error v/s Budget for data set COIL2

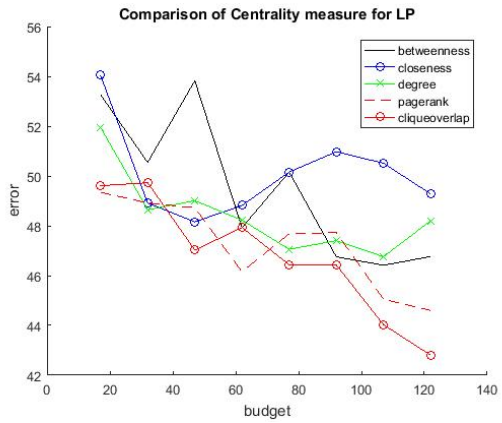

Fig. 13. Error v/s Budget for data set BCI

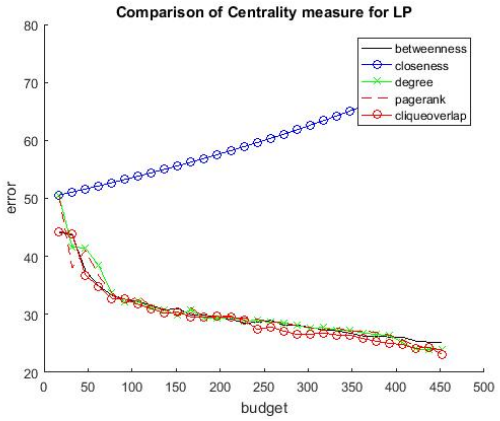

Fig. 14. Error v/s Budget for data set g241c

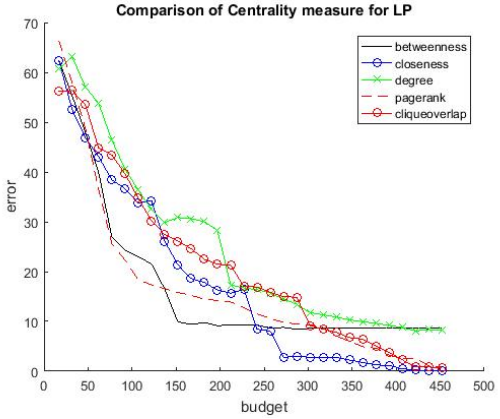

Fig. 15. Error v/s Budget for data set COIL

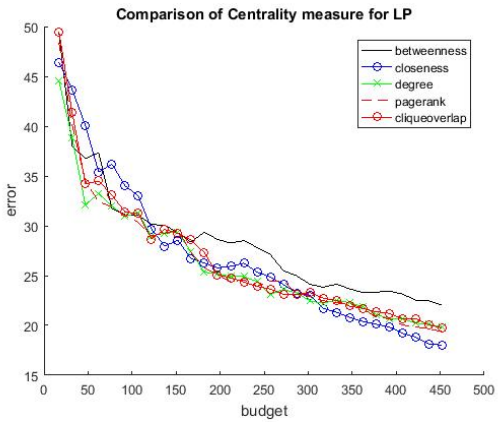

Fig. 16. Error v/s Budget for data set g241n

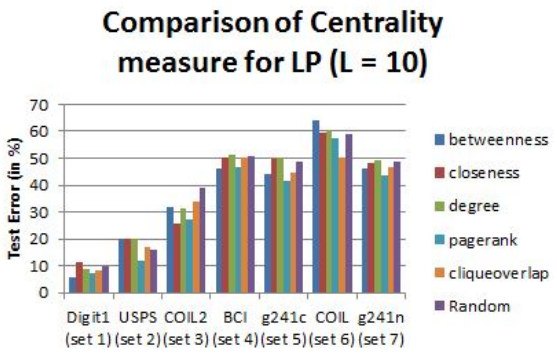

Fig. 17. Error v/s Budget for data set BCL

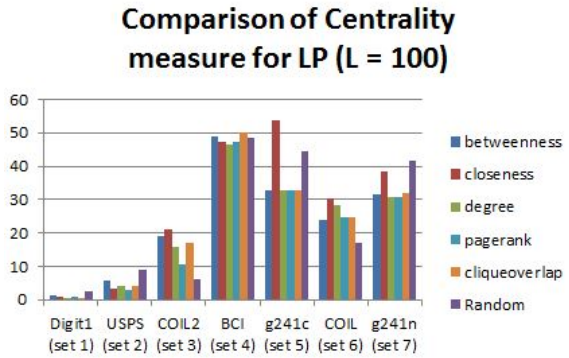

Fig. 18. Error v/s Budget for data set BCL

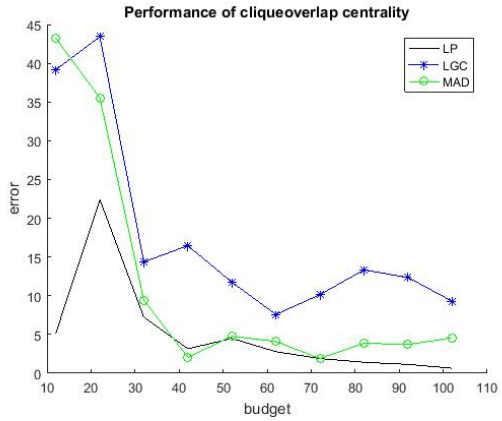

Fig. 19. Error v/s Budget for data set digit1

V. Conclusions

In our results we have included a random metric while comparing with other uncertainty metric. This random metric in active learning represents that we are selecting instances randomly and labelling them, i.e we have not taken care of informativeness of instance. As we can see that error thus obtained after executing G-ssl is more if we select instances randomly without taking care of informativeness of instance. Thus Active learning is actually helping in preparing most informative labelled set and unlabelled set.

Also we have observed that for calculating uncertainty metric out of entropy, least-confidence and margin, least-confidence gives us least error hence it can be concluded that least-confidence helps us in choosing best informative instance if only uncertainty is considered.

Now to consider instance correlation with uncertainty we use least-confidence with various centrality measures. We observe that clique-overlap centrality is most robust and stable of all othe centrality measures. Hence combination of Least-confidence and Cliqueoverlap helps us in choosing optimal informative instances.

We have also compared the error produced by MAD, Lable-propagation and LGC using the set $\{D_L\ D_U\}$ produced by Active learning which use least-confidence and Cliqueoverlap for utility metric. We Have observed that Lable-propogation and MAD gives the least error.

References

[1] M. R. Faghani and U. T. Nguyen, "A study of xss worm propagation and detection mechanisms in online social networks," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 11, pp. 1815–1826, 2013.

[2] Y. Fu, X. Zhu, and B. Li, "A survey on instance selection for active learning," *Knowledge and information systems*, vol. 35, no. 2, pp. 249–283, 2013.

[3] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[4] R. Motta, A. de Andrade Lopes, and M. C. F. de Oliveira, "Centrality measures from complex networks in active learning," in *Discovery Science*. Springer, 2009, pp. 184–196.

[5] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in *AAAI*, vol. 5, 2005, pp. 746–751.

[6] L. Bottou and P. Gallinari, *A framework for the cooperation of learning algorithms*. Citeseer, 1991.

[7] X. Zhu, Z. Ghahramani, J. Lafferty *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, vol. 3, 2003, pp. 912–919.

[8] P. P. Talukdar and K. Crammer, "New regularized algorithms for transductive learning," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2009, pp. 442–457.