Detailed Proposal for Active Learning Algorithm

Haohan Wang, Varuni Gang, Yucong Yang

November 18, 2013

1 Goal

We try to search for active positive regions predicted to be enriched P53 cancer rescue mutants, with half way point criterion, while trying to maximize F_α,
where $F_\alpha = (1+\alpha^2)\dfrac{precision * recall}{(1+\alpha^2)precision + recall}$
We propose a MIP-based active learning with a function of α , and we will dynamically select features.

2 Algorithm

```
*Initialization:*
    select $M$ features with information gain.
    randomly select $N_p$ positive instances and $N_n$ negative instances.
*Main Loop*
While:
    **1**. Select $n$ instances with $score_k$ for Instance $k$ is

        $$score_k = score_{classifier} + g(\alpha)w$$

    where $score_{classifier}$ is a score given by a traditional active learner
    and $g(\alpha)$ is a function of $\alpha$, it could be linear or exponential or something else
    **2**. Perform feature selection to select $M$ features with information gain
    **3**. Distinctly select $M$ top features with $\lambda M$ features from $feature\_set_{i-1}$ and $(1-\lambda)M$ top features from $feature\_selection_i$
    where $\lambda$ is constant in $[0, 1]$
*Terminate:*
    when active learner collected 50% of positive instances
```

3 Thinking Behind the Algorithm

1. For step 1 in our algorithm, we believe the function in Samuel et al paper gives a stable weight for predicted positive instance, which may only work for what they report, but may not work for other goal function. We smooth this weight by a parameter based on our goal function balancing *precision* and *recall*, so that this active learning method can give good results no matter what the goal in reality, hopefully.
2. For step 2 in our algorithm, we believe that since our classifier is built on different data sets in each round, then informativeness of features regarding to classifier may be different. If we only perform one feature selection at the very beginning of training, these features may not stay informative as the classifier evolves. In order to solve this problem, we select the most informative features every round. However, in order to avoid overfitting in feature selection process, each time we consider part of the features in last round.

4 Main Reference

Danziger, S.A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G.W., Kaiser, P., and Lathrop, R.H. (2009) *Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning*, PLOS Computational Biology, 5(9)