Budgeted Batch Mode Active Learning with Generalized Cost and Utility Functions

Arvind Agarwal, Shashank Mujumdar, Nitin Gupta, Sameep Mehta

IBM Research India

Introduction

- An ideal active learning algorithm should select examples in a batch that are maximally useful and can be labeled within a budget.
- Traditional algorithms define utility of a single example rather than of a batch, without cost and budget considerations.
- In this work we propose:
  - A novel optimization formulation based on knapsack problem.
  - A novel utility function based on the facility location problem that takes care of point utility, region utility and sample diversity.
  - A novel cost function that considers that the labelling cost of an example depends on the previously labelled examples.

Framework

Maximize the utility while minimizing the cost.

$$\boxed{\begin{aligned}&\max \sum_{i=1}^{N} v_i \mathbf{x}_i\\ &\mathbf{x}_i \in \{0,1\}\\ &\sum_{i=1}^{N} \mathbf{x}_i w_i \leq B\end{aligned}}$$

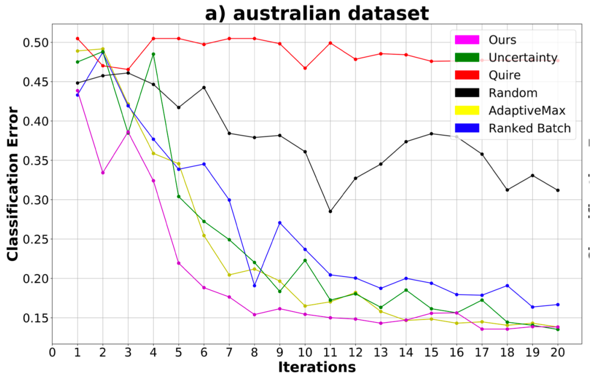
NP Hard so heuristics are needed

v_i = utility of the data point
w_i = labelling cost
B = Total budget

$$\boxed{\begin{aligned}\max_{\mathcal{X}} \quad & f_u(\mathcal{X})\\ & \mathcal{X} \subset \mathcal{U}\\ & f_c(\mathcal{X}) \leq B\end{aligned}}$$

$f_u(\mathcal{X})$ = Utility function operating on batch
$f_c(\mathcal{X})$ = labelling cost function operating no batch
B = Total budget

Representative Results (Australian Dataset)



Utility Function

- The utility function should:
  - Give higher weight to the important samples
  - Give higher weights to the region that is dense
  - Reduce redundancy or maximize the diversity
- The standard facility location problem takes care of all three characteristics

$$f_u(\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_k) = \sum_{i=1}^{N} e_i \max_{\mathbf{x}_k} f_s(\mathbf{x}_k, \mathbf{x}_i)$$

Cost Function

Cost is dependent on the previous labeled examples.

$$C(\mathbf{x}_i) = C_0(\mathbf{x}_i) + C(\mathbf{x}_i | \mathbf{x}_{i-1}, \mathbf{x}_{i-2} \dots)$$

Given such a cost function, one needs to find an optimal ordering which is given by Minimum Spanning tree.

Solution to the non-batch version is known to be NP-hard, however a practical solution is available based on dynamic programming.

$$K_{i,b} = \max(\underbrace{v_{i-1} + K_{i-1,b-w_{i-1}}}_{\text{When the next point is included}}, \underbrace{K_{i-1,w}}_{\text{When the next point NOT is included}})$$

Conclusion

- Proposed a novel and generic AL frame- work that selects the optimal batch within the budget constraint based on the given utility and cost functions.
- We also proposed a novel utility function based on the Facility Location problem.
- Proposed utility function has three important characteristics: (a) higher weights for important points, (b) higher weight for dense region, (c) Diversity of selected points.
- We also proposed a novel cost formulation, the optimal solution to which is the minimum spanning tree of the input batch.
- Experimental results on four datasets show that our approach outperforms the baseline algorithms especially in the initial iterations.