Optimal Subset Selection for Active Learning

Yifan Fu and **Xingquan Zhu**
QCIS Centre, Faculty of Eng. & Info. Technology, University of Technology, Sydney, NSW 2007, Australia
Yifan.Fu@student.uts.edu.au; xqzhu@it.uts.edu.au

Abstract

Active learning traditionally relies on instance based utility measures to rank and select instances for labeling, which may result in labeling redundancy. To address this issue, we explore instance utility from two dimensions: individual uncertainty and instance disparity, using a correlation matrix. The active learning is transformed to a semi-definite programming problem to select an optimal subset with maximum utility value. Experiments demonstrate the algorithm performance in comparison with baseline approaches.

Introduction

Active learning (Seung, Opper, and Sompolinsky 1992) reduces labeling cost by focusing on informative instances without compromising classifiers accuracy . Sample selection methods in active learning employ two types, (1) individual assessment based, and (2)data correlation based, of approaches. The former (Culotta and McCallum 2005) treats unlabeled instances as independent and identically distributed (I.I.D.) samples, without taking other samples into consideration. Data correlation based assessment (Nguyen and Smeulders 2004) uses sample correlations/distributions (*e.g.*, clustering) to select instances (*e.g.*, the centroid of each clusters) for labeling.

The key point of optimal subset selection is to ensure a selected labeling set containing mostly needed samples with minimum redundancy. When only considering instance uncertainty for labeling, a labeling set contains instances with the highest uncertainty values, whereas selected samples in the set may contain redundant knowledge so do not form an ideal candidate set, as shown in Fig.1(a). On the other hand, if we take instance uncertainty and disparity into consideration, we may form an optimal labeling set, where each samples in the set may not be the "most uncertain" ones, but together, they form an optimal labeling set. As shown in Fig. 1(b), the decision boundaries generated from six selected candidates are much closer to the genuine boundaries, compared to the approach in Fig.1(a). In this paper, we propose a new Active Learning paradigm using Optimal Subset Selection (ALOSS), which combines instance uncertainty and instance disparity to form a correlation matrix and select the instance subset with the maximum utility value. Such an instance selection problem is inherently an integer programming problem, which is NP-difficult but can be solved by using Semi-Definite Programming (*SDP*) (Goemans and Williamson 1995).

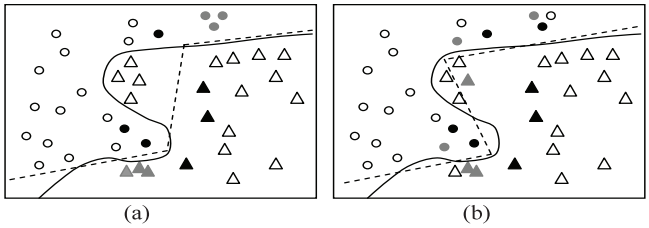

Figure 1: A toy example demonstrating labeling redundancy. Circles and triangles each denote one type of samples, with solid circles and triangles denoting labeled instances and the rest denoting unlabeled samples. The solid line denotes genuine decision boundaries and dashed lines denote decision boundaries learnt by learners.(a)samples selected by instance based assessment (b) samples selected by using optimal subset selection

Problem Definition & Algorithm Overview

Given a dataset $\mathcal{D}$ containing N instances $x_1, \cdots, x_N$, where samples are separated into a labeled subset $\mathcal{D}^L$ and an unlabeled subset $\mathcal{D}^U$, with $\mathcal{D} = \mathcal{D}^L \cup \mathcal{D}^U$ and $\mathcal{D}^L \cap \mathcal{D}^U = \emptyset$. The **aim** of optimal subset based active learning is to label a batch (*i.e.* a subset Δ) of instances, one batch at a time, from $\mathcal{D}^U$, such that when user requested number of instances are labeled, a classifier trained from $\mathcal{D}^L$ has the highest prediction accuracy in classifying test samples.

Assume a correlation matrix $\mathcal{M}$ is built to capture each single instance's uncertainty as well as the disparity between any two instances x_i and x_j, the above active learning **goal** can be regarded as the selection of an optimal subset of unlabeled samples Δ, such that the summation of instance uncertainty and disparity over Δ can reach the maximum. This problem can be formulated as a quadratic integer programming problem as follows,

$$\max_{\mathbf{e}} \mathbf{e}^T \mathcal{M} \mathbf{e}$$
$$s.t. \sum_{i, e_i \in \mathbf{e}} e_i = k \; ; \quad e_i \in \{0, 1\} \tag{1}$$

where $\mathbf{e}$ is an n-dimensional column vector and n is the size of unlabeled set $\mathcal{D}^U$. The constraint k defines the size of the subset for labeling, with $e_i = 1$ denoting that instance x_i is selected for labeling and $e_i = 0$ otherwise. Algorithm 1 describes the general process of our method.

Algorithm 1 ALOSS: Active Learning with Optimal Subset Selection

1: **while** *labeledSample* <budget **do**
2: construct a classifier ensemble with labeled training set D^L;
3: Build correlation Matrix $\mathcal{M}$
4: Apply optimal subset selection to $\mathcal{M}$ and select an optimal subset Δ with k instances;
5: *labeledSample* $\leftarrow$ *labeledSample* + k;
6: **end while**

Correlation Matrix Construction

To build a correlation matrix $\mathcal{M} \in \mathbb{R}^{n \times n}$, where n denotes the number of instances in the unlabeled set $\mathcal{D}^U$, we separate elements in $\mathcal{M}$ into two parts. More specifically, assume $\mathcal{U}_{i,i}$ defines the uncertainty of instance x_i and $\mathcal{I}_{i,j}, i \neq j$ defines the disparity between instances x_i and x_j, the correlation matrix $\mathcal{M}$ is constructed using Eq.(2)

$$\mathcal{M}_{i,j} = \begin{cases} \mathcal{U}_{i,j}, & \text{if } i = j \\ \mathcal{I}_{i,j}, & \text{if } i \neq j \end{cases} \tag{2}$$

Instance Uncertainty

To calculate uncertainty for each single instance x_i in $\mathcal{D}^U$, we build a classifier ensemble E with π heterogeneous members, $\hbar_1, \cdots, \hbar_\pi$ trained from labeled sample set $\mathcal{D}^L$. Denoting $\mathcal{H} \in \mathbb{R}^{\pi \times \pi}$ a normalized classifier weighting matrix where each element $\mathcal{H}_{i,j}$ denotes the agreement between classifiers $\hbar_i$ and $\hbar_j$ in classifying all instances in $\mathcal{D}^U$. After that, we apply each classifier $\hbar_j, j = 1, \cdots, \pi$, to x_i, and build a vector $\mathbf{u}_i$ with each element $u_{i,j}, j = 1, \cdots, \pi$, recording uncertainty of classifier $\hbar_j$ on instance x_i. As a result, we build an n by π matrix $\mathbf{u} \in \mathbb{R}^{n \times \pi}$, and calculate weighted instance uncertainty as follows.

$$\mathcal{U} = \mathbf{u} \times \mathcal{H} \times \mathbf{u}^T \tag{3}$$

Instance Disparity

We employ two types of distance measures, prediction distance and feature distance, to calculate disparity between each pair of instances x_i and x_j.

Prediction Distance ($\mathcal{P}$) captures prediction dissimilarity of a set of classifiers on two instances. For a pair of instance x_i and x_j, their prediction difference is accumulated prediction distance over all class labels and all classifiers.

Feature Distance ($\mathcal{F}$) captures the disparity of a pair of instances by using their Euclidean distance.

Instance Disparity ($\mathcal{I}$): Because prediction distance ($\mathcal{P}$) and feature distance ($\mathcal{F}$) each denotes the difference between instances x_i vs. x_j from different perspectives, the final disparity between x_i and x_j is the product of the two distances as follows.

$$\mathcal{I}_{i,j} = \mathcal{P}_{i,j} \times \mathcal{F}_{i,j} \tag{4}$$

Optimal Subset Selection

Using correlation matrix $\mathcal{M}$, the objective function defined in Eq.(1) is to select a k instance subset such that the summation of all instances' uncertainty and their disparities is the maximum among all alternative subsets with the same size. This problem is NP-difficult. We use *SDP* approximation algorithm "Max cut with size k" (MC-k) problem (Goemans and Williamson 1995) to solve this maximization problem with polynomical complexity.

Experimental Results

In Fig. 2, we compare accuracy between ALOSS and several baseline approaches on "auto" dataset, where "Entropy" denotes entropy-based uncertainty sampling (which has the same batch size Δ as ALOSS) and "SIB" represents Single Instance Batch active learning (which repeats for each single instance). From Fig. 2, it is clear that Entropy is the least effective algorithm because it does not take labeling redundancy into consideration. SIB can reduce labeling redundancy to some extent (by labeling instance one at a time), but it still cannot guarantee an optimal labeling set. By combing instance uncertainty and disparity to select optimal subsets, ALOSS outperforms all baseline approaches.

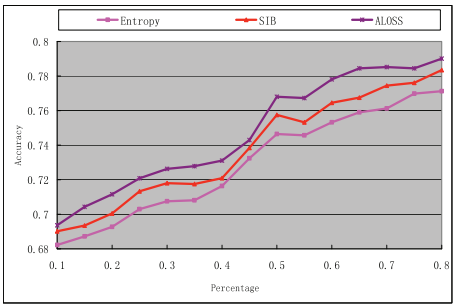

Figure 2: Algorithms accuracy comparison between ALOSS and baseline methods *w.r.t.* different percentages of labeled data

Conclusion

We proposed a new active learning paradigm using optimal subset selection (ALOSS), which considers uncertainty and disparity of an instances subset to reduce labeling redundancy and achieve good performance.

Acknowledgement

This research is supported by Australian Research Council Future Fellowship under Grant No. FT100100971.

References

Culotta, A., and McCallum, A. 2005. Reducing labeling effort for stuctured prediction tasks. In *Proc. of AAAI*, 746–751.

Goemans, M., and Williamson, D. 1995. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite prgramming. *J. of ACM* 42.

Nguyen, H., and Smeulders, A. 2004. Active learning using pre-clustering. In *Proc. of ICML*, 623–630.

Seung, H.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proc. of COLT*, 287–294.