Combining Link and Content for Collective Active Learning

Lixin Shi, Yuhang Zhao, and Jie Tang
Dept. of Computer Science and Technology, Tsinghua University
Beijing 100084, China
{shilixinhere, zhaoyh630}@gmail.com, tangjie@keg.cs.tsinghua.edu.cn

ABSTRACT

In this paper, we study a novel problem *Collective Active Learning*, in which we aim to select a batch set of "informative" instances from a networking data set to query the user in order to improve the accuracy of the learned classification model. We perform a theoretical investigation of the problem and present three criteria (i.e., minimum redundancy, maximum uncertainty and maximum impact) to quantify the informativeness of a set of selected instances. We define an objective function based on the three criteria and present an efficient algorithm to optimize the objective function with a bounded approximation rate. Experimental results on a real-world data sets demonstrate the effectiveness of our proposed approach.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation

Keywords

collective active learning, link, document classification

1. INTRODUCTION

Machine learning algorithms suffer from insufficiently labeled training data. The goal of active learning is, as usual to construct an accurate classifier, but also to minimize the number of labeled instances by actively selecting a few number of instances to query the user. Traditionally, this problem is usually addressed in a single mode, i.e., the active learning algorithm queries the user k times, with each time querying one instance for its label. Following this thread, considerable research has been conducted on how to select the best example to query in each time [7, 11].

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

Recently, there has seen a new direction of machine learning field, that is how to learn an accurate model to classify networking/graphical data, e.g., the linked Web pages. Quite a few models have been proposed such as Conditional Random Fields [2], Continuous Bayesian network [6], Collective Learning [1], and Semi-supervised Learning over graphs [10]. A few works also try to combine the networking information under single mode active learning framework [3, 8, 11]. However, two important issues have been largely ignored by most existing works. First, almost all learning algorithms for the networking data are computationally intensive. Suppose a machine needs to query the user k times, when the user inputs a label for the queried instance, she/he may have to wait for a very long time for the next query, which is obviously undesirable. Second, these methods are in a single mode, the selected instances in different iterations may have a undesirable information overlap.

Ideally, we hope that an algorithm can actively select a set of instances with minimal redundancy to query the users in a batch mode, which we refer to as the *collective active learning* (CAL) problem for networking data. The problem posts several unique challenges. First, as the optimization problem of selecting the most informative instances is NP-hard, it is unclear how to formulate the problem in a principled framework. Second, to design criteria to quantitatively measure the informativeness of the data is not an easy task. Third, the active learning algorithm should be efficient, in particular considering the rapidly increasing scale of the networking data on the Web.

To this end, we formally formulate the problem and propose a general framework for collective active learning. Specifically, we propose three criteria to respectively capture the maximum uncertainty, maximum impact, and minimum redundancy (which will be explained in section 3.1). We design an objective function based on the criteria and further propose an efficient algorithm to solve the objective function. A theoretical analysis for the approximation rate of the algorithm is presented. We conduct experiments on a real-world data set to validate the effectiveness and efficiency of the proposed approach. Experimental results show that our approach clearly outperforms (+6%) the baseline methods of single mode active learning and batch mode active learning on linked data sets.

2. PRELIMINARIES

The collective active learning problem can be defined as follows: given an (un)directed graph $\mathcal{G} = (V, E)$, where V indicates a set of data points and $E \subset V \times V$ represents a

set of edges between the data points. For example, in a social network, the edge can represent the friendship between users; while in the citation network, the edge presents the citation relationship between papers. Suppose there are n unlabeled data $\mathcal{U} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$, where $\mathcal{U} \subset V$ and l labeled data $\mathcal{L} = \{(\boldsymbol{x}_{n+1}, y_{n+1}), \cdots, (\boldsymbol{x}_{n+l}, y_{n+l})\}$, where $\mathcal{L} \subset V$. In most circumstances, we have $l \ll n$. Further let $\boldsymbol{x}_i$ denote the observation vector, e.g., it can be the feature vector in most applications. Without loss of the generality, we associate each data point with a binary classification label, $y_i \in \{0, 1\}$.

A general classification problem is to learn a mapping function f from the labeled data points $\mathcal{L}$ to predict the labels of the unlabeled data points $\mathcal{U}$. However, as labeling is always tedious and time consuming, the labeled data points are usually insufficient. The problem of collective active learning is to select $k \ll n$ unlabeled data points, i.e., $S \in \mathcal{U}$ with $|S| = k$, to query their labels, in order to improve the prediction accuracy of the learned function f. The goal is to maximize the improvement on the accuracy by querying the k data points. Formally, we can define the following objective function:

$$Q : 2^{\mathcal{U}} \to \mathbb{R}, \text{ in } (\mathcal{U}, \mathcal{L}, G)$$

And the goal is to select S to maximize the function Q:

$$S = \underset{S \subseteq \mathcal{U}, |S| \leq k}{\operatorname{argmax}} \{Q(S)\}$$

In this way, the collective active learning problem can be also considered as a set function optimization problem. The following task is how to instantiate the objective function $Q(S)$ and how to efficiently solve the function $Q(S)$.

Three Criteria In our collective active learning problem, one key challenge is how to measure the informativeness of a set of selected instances. In this work, we propose three criteria to measure the informativeness of instances.

- *Maximum Uncertainty*: We are more willing to choose samples which tend to be more uncertain. One intuitive method for binary classification is to choose the instances whose posterior probabilities of being positive is nearest to 0.5.

- *Maximum Impact*: Selected samples should have the maximum impact on the unknown instances. If a sample is isolated in the sample space, for example, an outlier, it should be given a low priority to be selected. Note the impact is two-fold: the similarity between the two feature vectors and the connecting edges both imply "impact" information.

- *Minimum Redundancy*: The samples in the selected set should be diversely distributed over the space. In other words, it minimizes the information overlap between the selected samples.

3. THE PROPOSED APPROACH

3.1 Objective Function

Based on the defined criteria, we give an instantiation of the objective function. This is just a possible meethod, but not the only way to instantiate the objective function. Basically, the objective function is defined as a linear combination of the two terms, i.e., $C(S)$ and $H(S)$:

$$Q(S) = \alpha C(S) + (1 - \alpha) H(S), \quad 0 \leq \alpha \leq 1 \tag{1}$$

where $H(S)$ corresponds to the maximum uncertainty and $C(S)$ corresponds to the maximum impact.

Maximum Uncertainty We use entropy to measure the uncertainty of selected samples. Joint entropy is very hard to compute, so we use the summation of entropies over single data points. In summary, the maximum uncertainty part is defined as the $H(S)$ function in $Q(S)$:

$$H(S) = \sum_{i \in S} H(i) = \sum_{i \in S} f_i \log \frac{1}{f_i} + (1 - f_i) \log \frac{1}{1 - f_i} \tag{2}$$

Maximum Impact The motivation of our maximum impact measurement comes from the classical nearest neighborhood classifier. The classifier classifies data point $\boldsymbol{x}_i$ into the same class with labeled data point $\boldsymbol{x}_j$ which has the highest impact on $\boldsymbol{x}_i$:

$$Class(\boldsymbol{x}_i) = y_j, \quad j = \underset{j \in L}{\operatorname{argmax}} w_{ij}$$

From the view of Nearest Neighbor classifier, the classification result is more guaranteed if the impact is higher. That gives a direct motivation on the maximum impact measurement: to maximize the impact on a single unlabeled data point $\boldsymbol{x}_i$, we can choose the data points with the maximum impact over $\boldsymbol{x}_i$ from the candidates. So we can have a weighted function of summations over all these maximum values to measure the impact:

$$C(S) = \sum_{i \in U} s_i \max_{j \in L \cup S} w_{ij} \tag{3}$$

where s_i serves as a weight factor when counting the impact over points in the unlabeled data set. It has no problem to choose $s_i = 1$ for all unlabeled data points, but there may be better choices. We suggest to use entropy as the weight. Specifically,

$$s_i = H(i) = f_i \log \frac{1}{f_i} + (1 - f_i) \log \frac{1}{1 - f_i}$$

The point is that the use of entropy information here does not overlap with the entropy in $H(S)$: different examples are checked by C(S) and H(S) in terms of entropy. To achieve a higher flexibility, we can introduce a balancing factor β to give a strengthened definition of $C(S)$ as(for $i = j$, $w_{ij} = 1$):

$$C(S) = \sum_{i \in U} (H(i))^{\beta} \left(\max_{j \in L \cup S} w_{ij} \right)^{1 - \beta} \tag{4}$$

Minimum Redundancy In equation 1, we do not have a term explicitly demonstrating the redundancy over the selected set. In this section, we'll prove that the minimum redundancy criterion has already been implicitly satisfied in the definition of $Q(S)$.

The following is an explanation why maximizing $Q(S)$ will also minimize the diversity. Specifically, $C(S)$ is closely related to redundancy. Given a data point $i \in U - S$, let us define the dominant point $dp(i)$ as

$$dp(i) = \max_{j \in S \cup L} w_{ij} \tag{5}$$

The maximization of $Q(S)$ will cause the dominant points get diversely distributed. If two dominant points in S are very close to each other, it is likely that they may have similar impact on other points, thus removing one of them will not let $C(S)$ decrease much. In other words, if we already have vertex i in the selected set S, we'll not choose another vertex j similar to i in the future, because the improvement on $Q(S)$ is little.

3.2 Combine Link Information in W

It is flexible in our framework that link information is naturally integrated into the definition of similarity matrix, by extending it using a similar method as page rank. That is reasonable because edges indicate the similarity and impact between the two ends. It is introduced that the similarity matrix W is used as transformation probability matrix in random walk[11]. Page rank is a way to introduce the graph structure into the transformation matrix. Generally speaking, under the page rank model, a particle may transit in one of the following cases:

- It may transit by edges. The particle will transit with equal or weighted probability to each of the neighboring vertices.
- It may randomly jump to any vertex. The probability is proportional to similarity in feature space.

Suppose there's a well defined similarity matrix W which measures the impact solely in feature vector space, now we want to integrate link information into it, to get a new definition $\widehat{W}$.

$$\hat{w}_{ij} = \epsilon \frac{1}{d_i} I(i,j) + (1-\epsilon)\frac{w_{ij}}{\sum_k w_{ik}}$$

where $0 \leq \epsilon \leq 1$, $I(i,j)$ is an indicator function whether there is an edge between point i and j:

$$I(i,j) = \begin{cases} 1 & (i,j) \in E \\ 0 & (i,j) \notin E \end{cases}$$

and d_i is the degree of i, $d_i = \sum_{(i,j)\in E} 1$.

3.3 EFFICIENT ALGORITHM

It can be proved that our algorithm is montonic submodular. There have been many works on finding good approximation algorithms for monotone submodular function optimization. For simplicity and efficiency, we'll use the greedy algorithm [5]. Algorithm 1 shows a structure of this algorithm. The outline of the algorithm is repeatedly enlarging set S by a new point v, such that $Q(S\cup\{v\}) - Q(S)$ is maximized. This algorithm have a guaranteed approximation rate $1 - \frac{1}{e}$.

Algorithm 1 Maximize $Q(S)$

Input: U, L, G, $\mathbf{x}$, $\mathbf{y}$, k
Output: $S, |S| = k$

```
Calculate probability vector f
for v ∈ U calculate H(v)
initialize S ← ∅
while |S| < k do
    for v ∈ U − S calculate C(v) = C(S ∪ {v}) − C(v)
    find v ∈ U − S to maximize αC(v) + (1 − α)H(v)
    update S ← S ∪ {v}
end while
```

4. EXPERIMENTS

4.1 Data Set

The experiments are performed in text categorization data sets with citation information. We test the proposed method on the following three data sets, which are the most widely used data sets for text classification with link information:

Cora Data Set [9] contains 2708 scientific publications, and they are classified into seven fields, which are Case Based, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning, and Theory. After stemming and removing stop words, we are left with a vocabulary size of 1433 unique words, all of which appear at least 10 times in the documents. There are 5429 citations between the documents. We construct a binary classification problem by combining the first 4 classes into a category, and the others as another category.

Citeseer Data Set [9] contains 3312 publications, labeled into 6 classes: Agents, AI, DB, IR, ML, and HCI. There are 3703 unique words after processing, and the number of citations is 4732. In the same way as in Cora data set, we construct it into the binary classification by grouping some classes into a category.

WebKB Data Set [9] contains web pages from four computer science departments, categorized into five topics: course, faculty, student, project, and staff. The webKB data set contains 877 data points and 1703 unique words. There are 2868 total links between these pages. We construct it to binary classification by letting class "course" and "project" be one class and the others be another class.

4.2 Baselines

For the linked data set, we use the following methods as baselines:

Random selects the samples set randomly, giving each unlabeled point equal probability to be selected.

Most uncertainty selects the set with the largest entropy. Specifically, it is the function when $\alpha = 0$.

Active Learning using Gaussian Fields is an approach suggested by [11] based on a semi-supervised learning framework using Gaussian fields and harmonic functions. It is single-mode, so we run this algorithm k times to select a set of size k. Note that in this framework, the link information can be similarly introduced using the proposed method. We will utilize the link information in the tests.

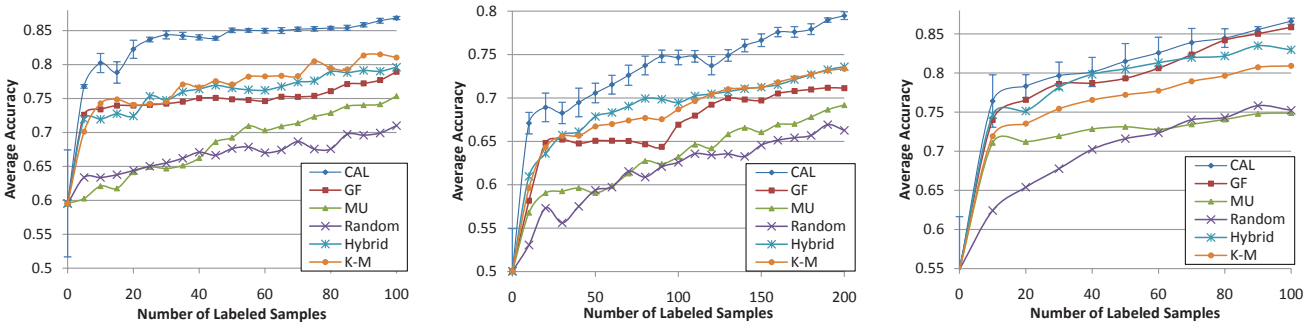
Hybrid is suggested by [3]. It asks for uncertainty approach and two graphical metrics (betweenness and cluster-finding) to find a selected set, and using empirical risk to pick the best set among the union of the data points selected by the three strategies.

k-means is suggested by [8]. In the article some active inference methods are compared with each other and k-means is found to be the best one among them. Here we employ the same strategy for active learning, that is, we find vertices using k-means as the labeled set and then train the classifier.

For the purpose of simplicity, we will use **Random, MU, GF, Hybrid, K-M** short for the methods above respectively. We refer to our model as **CAL** (Collective Active Learning).

4.3 Results

We set the α parameter to be 0.5, meaning each criterion



(a) Accuracy in Cora Data Set (b) Accuracy in Citeseer Data Set (c) Accuracy in WebKB Data Set

Figure 1: Tests on Three Data Sets with Links

are roughly equally balanced in our experiment. For the cora data set, we randomly pick 5 points as the initially labeled set $\mathcal{L}$; for the other two data sets, this number is 10 due to the size of the data set and the learning difficulty. For the same reason, we define the batch size $k = 5, 10, 5$ for Cora, Citeseer and WebKB data sets respectively. For each data set, the batch mode active learning and collective active learning methods first select k samples based on $\mathcal{L}$, and then repeatedly select k samples based on the union of initialized labeled and selected samples; the single mode active learning iteratively select k samples, and repeatedly select and update the model as the batch mode methods do. After the selecting process, we learned the prediction based on the samples selected by different active learning methods using the same semi-supervised learning method for fair. Here we use the famous NetKit-SRL toolkit[4] for learning in networked data set. For each data set, we run the experiment 30 times with different initially labeled sets, and both the average and variance of the accuracy is used for final evaluation.

Figure 1 shows the results on each of the data set. Due to space limitations, we only draw the variance of proposed method in the figure. We can show from the results that maximum entropy does not have a good performance over all the data sets. For all the three data sets, our method outperforms the strategies based on graph metrics: the hybrid method and the k-means method. The Gaussian random field based method's performance in the three results was somewhat erratic: it does not perform well in Cora and Citeseer data set but in the webKB data set, the accuracy of it is very near to the proposed method. In webKB data set, the gap between random selection and these methods are not as high as other data sets. It is probably because it is not so easy in this web-linked data set to perform collective active learning. Also, from the view of variance, our methods have average variances of 0.005, 0.009, 0.01 on the Cora, Citesser and WebKB data set, which are much smaller than the other methods. Generally speaking, from the experiment results of the three linked data set, we can conclude that our method is the best, or at least close to the best.

5. CONCLUSION

In this paper, we present a novel framework for collective active learning, which utilize both link and content information. An objective function is defined based on three proposed criteria. Experiments on a real-world data set shows that our approach outperforms other state-of-the-art methods in both linked and regular data sets. Although our model concentrates on the binary classification problem, it can be easily extended to the multi-class classification problem.

6. *ACKNOWLEDGMENTS

The work is supported by the Natural Science Foundation of China (No. 60703059), Chinese National Key Foundation Research (No. 60933013), National High-tech R&D Program (No. 2009AA01Z138).

7. REFERENCES

[1] D. Jensen, J. Neville, and B. Gallagher. Why collective inference improves relational classification. In *KDD '04*, pages 593–598, 2004.

[2] J. Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01*, pages 282–289, 2001.

[3] S. A. Macskassy. Using graph-based metrics with empirical risk minimization to speed up active learning on networked data. In *KDD '09*, pages 597–606, 2009.

[4] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983, 2007.

[5] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[6] U. Nodelman, C.R. Shelton, and D. Koller. Learning continuous time bayesian networks. In *UAI '03*, pages 451–458, 2003.

[7] S. Rajan, D. Yankov, S. J. Gaffney, and A. Ratnaparkhi. A large-scale active learning system for topical categorization on the web. In *WWW '10*, pages 791–800, 2010.

[8] M. J. Rattigan, M. Maier, and D. JJensen. Exploiting network structure for active inference in collective classification. In *ICDM Workshop '07*, pages 429–434, 2007.

[9] P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[10] X. Zhu. *Semi-Supervised Learning with Graphs*. PhD thesis, Carnegie Mellon University, 2005. CMU-LTI-05-192.

[11] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML Workshop '03*, pages 58–65, 2003.