УДК 681.3

E. A. Engel

AN ANALYSIS OF INTELLIGENT METHODS AND ALGORITHMS FOR UNLABELED DATA PROCESSING*

Intelligent algorithms and method is well-suited to many problems in data processing, where unlabeled data may be abundant. We survey previously used selection strategies for intelligent model, and propose two novel algorithms to address their shortcomings, focus on Active Learning (AL). While has already been shown to markedly reduce the annotation efforts for many sequence labeling tasks compared to random selection, AL remains unconcerned about the internal structure of the selected sequences (typically, sentences). We propose a semi-supervised AL approach for sequence labeling.

Keywords: intelligent algorithms and methods, data processing, active learning.

Modeling can have a number of objectives, including understanding or explaining data, developing scientific theories, and making predictions. We focus in this paper on predictive modeling. The goal is to predict an outcome y given a number of predictor variables $x = [x_1, x_2, \dots x_n]$, also called features, attributes, or factors. For instance, in handwriting recognition, the predictor variables may indicate the presence of features such as loops or segments oriented in a certain direction, and the target variable or "label" can be a letter or a word to be recognized. In chemo-informatics, the predictors are descriptors of the molecular shape and the target indicates e.g., the activity of the molecule against a given disease. In text processing, the predictors may be simple frequencies of words and the target could be a document category such as "politics", "sports", "computers", etc.

Producing predictive models in the "supervised learning" setting requires a "training set" of labeled examples, i.e., examples x for which the target value or "label" y is known. Those are used to train the predictive model, which is then evaluated with new examples (test set) to estimate its "generalization performance".

In many applications, including handwriting recognition, chemo-informatics, and text processing, large amounts of unlabeled data are available at low cost (x only is known), but labeling examples (using a human expert to find the corresponding y value) is tedious and expensive. Hence there is a benefit to either use unlabeled data to improve the model in a semi-supervised learning algorithm, or to sample efficiently x-space to use human expertise for labeling only the most informative examples.

In the regular machine learning setting (passive learning), a batch of training pairs $\{x, y\}$ is made readily available (training set). The learning machine may be used to select the examples, which look most promising to improve the predictive model. There exist several variants of active learning:

– pool-based active learning: a large pool of examples x is made available from the onset of training;

– stream-based active learning: examples are made available continuously;

– de novo query synthesis: the learner can make up values of x.

Of the variants of active learning considered, pooled-based active learning is tremendously important in today's machine learning and data mining applications, because of the availability of large amounts of unlabeled data in many domains, including pattern recognition (handwriting, speech, airborne or satellite images, etc.), text processing (internet documents, archives), chemo-informatics (untested molecules from combinatorial chemistry), and marketing (large customer databases).

Stream-based active learning is also important when sensor data is continuously available and data cannot be easily stored. However, it is more difficult to evaluate. It is reasonable to assume that several of the techniques developed for pooled-based active learning will also be applicable to stream-based active learning. The problem of de-novo queries is conceptually rather different because it involves human interventions on the system that may disrupt its normal functioning (interventions or manipulations).

A number of query strategies with various criteria of optimality have been devised. Perhaps the simplest and most commonly used query strategy is uncertainty sampling [1]. In this framework, an active learner queries the instances that it can label with least confidence. This of course requires the use of a model that is capable of assessing prediction uncertainty, such as a logistic model for binary classification problems. Another general active learning framework queries the labels of the instances that would impart the greatest change in the current model (expected model change), if we knew the labels. Since discriminative probabilistic models are usually trained with gradient-based optimization, the "change" imparted can be measured by the magnitude of the gradient [2].

A more theoretically motivated query strategy is query-by-committee (QBC) [3]. The QBC approach involves maintaining a committee of models, which are all trained on the current set of labeled samples, but represent competing hypotheses. Each committee member votes on the labels of query candidates and the query considered most informative is the one on which they disagree most.

*Работа выполнена при финансовой поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг.

It can be shown that this is the query that potentially gives the largest reduction in the space of hypotheses (models) consistent with the current training dataset (version space). A related approach is Bayesian active learning. In the Bayesian setting, a prior over the space of hypotheses gets revised into a posterior after seeing data. Bayesian active learning algorithms, for instance [4], maximize the expected Kullback-Leibler divergence between the revised posterior distribution (after learning with the new queried example) and the current posterior distribution given the data already seen. Hence this can be seen both as an extension of the expected model change framework for a Bayesian committee and a probabilistic reduction of hypothesis space.

A more direct criterion of optimality seeks queries that are expected to produce the greatest reduction in generalization error (expected error reduction). The first statistical analyses of active learning proposed in [5], demonstrating how to synthesize queries that minimize the learner's future error by minimizing its variance. However, their approach applies only to regression tasks and synthesizes queries de novo. Another more direct, but very computationally expensive approach is to tentatively add to the training set all possible candidate queries with one of the opposite label and estimate how much generalization error reduction would result by adding it to the training set [6].

It has been suggested that uncertainty sampling and QBC strategies are prone to querying outliers and therefore are not robust. The information density framework [7] addresses that problem by calling informative instances that are not only uncertain, but representative of the input distribution.

Active Learning. Sequence labeling is the task of mapping an ordered list of inputs to a sequence of output tags. It has many practical applications in natural language processing such as named entity recognition, part-of-speech tagging, shallow parsing, and text chunking. Another potential application, which is investigated in this study, is the subphrase generation problem. The goal of subphrase generation in query processing is to find subphrases in a query that maximally preserve the user's intent. Unlike the classification of record based data, sequence labeling depends not only on the features extracted from the input equence but also on its previous output tags. Many algorithms have been proposed in the literature to address this problem, including Conditional Random Field [8], Hidden Markov Model [9] and Maximum Entropy Markov Model [10].

A known problem in supervised learning tasks such as sequence labeling is the difficulty of acquiring labeled examples. The size of training data available is often limited because labeling examples can be very expensive. Labeling a sequence is also more challenging because the output tag depends on both the input and previous output tags. As a result, the tags of a sequence must be determined as a whole, rather than individually for each input element. Active learning may help to address this problem by selecting a small subset of examples for labeling from the large pool of unlabeled sequences available. By selecting the most informative examples, active learning can significantly reduce the required size of training data while maintaining comparable level of performance. However, the definition of «informative» varies for different algorithms and applications. One commonly used method is to select examples with largest uncertainties. In this paper, we treat each sequence as a whole for labeling and propose two strategies to measure the uncertainty of sequences under the Neural Network framework, referred as simple uncertainty (SU) and most-possible-constraint-violation method (MPSV).

Sequence Labeling Problem. Sequence labeling is a common problem with many applications in many areas such as named entity recognition [11], POS tagging [12], text chunking [12], etc. Definition 1 and Definition 2 give the formal definition of sequence and sequence labeling problem.

Definition 1 [Sequence]: A sequence x is an ordered list of elements $x = (x^1, x^2, ..., x^t)$.

Definition 2 [Sequence Labeling]: Given a sequence of inputs x, the sequence labeling problem is trying to label it with a sequence of tags $y = (y^1, y^2, ... , y^t)$, where each tag y^i belongs to a tag set D with $|D|$ tags.

One simple way to solve the sequence labeling problem is to use traditional classification algorithm such as Neural Network, which treats each element in the sequence as one example. However, it requires the features extracted only depend on the inputs x, which is not true in sequence labeling problem. The features extracted for sequence labeling not only depends on the inputs x, but also depends on the outputs y. The feature vector for a sequence (x, y) is represented as a joint feature mapping vector $\varphi(x, y)$. The definition of φ depends on the nature of different applications. One example feature for the subphrase matching problem would be "previous word is dropped → current word is kept", which represents the transition from previous tag "0" to current tag "1".

Now assume that we have a training sequence set $X = \{x_1, x_2, ..., x_n\}$ with its corresponding tag sequence set $Y = \{y_1, y_2, ..., y_n\}$. We are interested in learning a mapping function $f : X \rightarrow Y$. Instead of learning f directly, the strategy is to transform the problem into learning a discrimination function F over the joint mapping of input and output:

$$X \times Y \rightarrow R.$$

Given a test sequence x, its prediction is achieved by maximizing F over the response variable. The generalized form of the hypotheses f becomes

$$f(x, w) = \arg\max_{y \in Y} F(x, y; w)\,, \qquad (1)$$

where w is the parameters to be learned. Using the joint feature vector $\varphi(x, y)$, it can be further formulated as

$$f(x, w) = \arg\max_{y \in Y} F(\varphi(x, y); w). \qquad (2)$$

Note that many existing methods for sequence labeling problem can be explained in the above framework. For example, the function form F that are maximized in the above prediction function represents the conditional probability $P(y|x)$ in conditional random filed [8], Hidden Markov Models [9], Maximum Entropy Markov Models [10] and Modified Neural Network [13].

Active Learning By Modified Neural Network. From the previous work on active learning [14], measurement of uncertainty has played an important role in selecting the most valuable examples from a pool of unlabeled data. In the in the above framework, three methods have been proposed to measure the uncertainty of simple data, which are referred as simple margin (fig. 1), MaxMin margin and ratio margin.

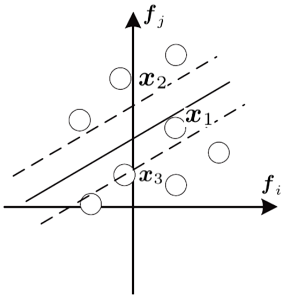

Fig. 1. Simple margin method will select unlabeled data x_1 for querying, which lies closest to the hyperplane

Simple margin measures the uncertainty of an simple example x by its distance to the hyperplane w calculated as

$$|w \cdot \varphi(x)|. \qquad (3)$$

As illustrated in fig. 1, examples lying closer to the hyperplane are assigned with larger uncertainty score. This is consistent with the intuition that examples close to the hyperplane are classified with lower confidence. These examples are considered as valuable examples since they have higher probability to be misclassified and thus more informative to be selected for further training.

However, labeling an element in a sequence by itself is almost infeasible in most sequence labeling applications because of the requirement for context information. In most situations we have to consider a whole sequence as an unit for uncertainty measurement and active selection. Given a pool of unlabeled sequences, $U = \{s_1, s_2, ..., s_m\}$, the goal of active learning in sequence labeling is to select the most valuable sequences from the pool. A straightforward way to measure the uncertainty of a sequence s is by its prediction score. The prediction score $w^T\ \varphi(s, y)$ measures the certainty of labeling test sequence s using the tag sequence y.

The simple uncertainty for sequence s is then calculated in Neural Network as:

$$UC(s) = \exp(-\max_{y \in Y} w^T\ \varphi(s, y)), \qquad (4)$$

which is based on the negative value of the prediction score given by formula (2). Note that the features in sequence labeling not only depend on the input sequence s, but also depends on the output y. Finally, the sequences with larger uncertainty are selected as valuable examples to add to the training set for further learning. We refer this method as simple uncertainty (SU) in this paper.

One drawback of the simple prediction score is its ignorance of the underlying score distribution among different classes and only use the maximum score as a measure of certainty. Here we propose another method which defines the uncertainty of a sequence x as

$$\mathrm{MPSV}(s) = \exp(-\max_{\substack{y_1, y_2 \in Y \\ y_1 \neq y_2}} (w^T \varphi(s, y_1) - w^T \varphi(s, y_2))), \qquad (5)$$

which can be further formulated as

$$\mathrm{MPSV}(s) = \exp(\min_{y \in Y} w^T\ \varphi(s, y) - \max_{y \in Y} w^T\ \varphi(s, y)). \qquad (6)$$

We measure the uncertainty of an sequence s as the difference between the minimum prediction score and the maximum prediction score, which is actually the most possible violated constraint for a sequence s that can be added into the optimization problem.

We refer this method as the most-possible-constraint-violation method (MPSV) in this paper. The two methods SU and MPSV proposed here are used to calculate the uncertainty for each test sequence s. The test sequences with maximum uncertainty score are selected as the most informative sequences. These sequences are submitted to the labeler to query for labeling and further added into the training set.

Experiment Result. We applied our algorithm to three data sets in our experiment. The first two data sets come from named entity recognition shared task of CoNLL-2002 [11]. One is Spanish data (ESP), which is a collection of news wire articles made available by the Spanish EFE News Agency. Another is Dutch data NED, which consist of four editions of the Belgian newspaper "DeMorgen" of 2000. The task is to label each word in the sentence using some predefined entity tags such as person names (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC) with a B ahead of them denoting the first item of a phrase and an I any non-initial word. The third data we are using is collected from the query subphrase matching project (QSPM) of Yahoo Sponsor Search. Given a query by a typical search engine user, the goal is to generate subphrases that preserve the user intent as well as match the bidded terms submitted by advertisers. There are two tags: "KEEP" ("1") and "DROP" ("0") for each position.

For each position in the sequence, we extract its context features such as "current word is", "previous word is", "next word is" and so on. We also used tag transition features such as "previous tag to current tag". Some word features such as prefix and suffix are also used based on the language of the data such as "th" for English data. We did not employ any feature selection methods in our experiments. For the DER data, the Part-Of-Speech tags are also utilized as grammatical features.

In this experiment,we compare themost-possible-constraint-violationsmethod (MPSV) and simple uncertainty(SU) method with the random method. To alleviate the length problem in sequence active learning, we select a subset of sequences from the training data, which has the same length. For each data set, we run four experiments, each on a different length selected from the training data. For NED data, we select all the sequences with length 12, 13, 14 and 15 in each experiment. For ESP data, we select all the sequences with length 42, 43, 44, 45 in each experiment. For the QSPM data, we select all the sequences with length 3, 4, 5, 6. For the NED and MPSV data set, we select 400 sequences at each length. The first 10 are used for initial training. The pool of the remaining 390 sequences is for active selection. Each time we select 15 sequences and the result is reported as the average error rate of different length. For the QSPM data, we select 1930 sequences at each length. The first 10 sequences are used for initial training. The pool of the remaining 1920 sequences is for active selection. Each time we select 60 sequences and the result is reported as the average error rate of different length on the test set.

Fig. 2 shows the results for the three methods in the three data sets ESP, NED and QSPM. The *x*-axis denotes the number of unlabeled sequences selected to query for labeling. The *y*-axis represents the average error rate, which is calculated in the word level as follows:

$$\mathrm{ErrorRate}_{\{\mathrm{WordLevel}\}} = = \frac{\text{Total number of correctly tagged words}}{\text{Total number of words}}. \qquad (7)$$

We observe from the fig. 2 that both MPSV and SU methods outperform random approach on all three data sets. Also MPSV performs better than SU, which means that MPSV is a better way to measure uncertainty for Modified Neural Network. Furthermore, the gap between the MPSV and other two methods seems very large when the number of selected sequences is small. It means that MPSV serves as a good criteria that only a small number of sequences are needed to get good performance. In this experiment, all the sequences are of the same length to compare three methods and we are aiming to select a predefined number of sequences.

In this paper, we have proposed two measurements of uncertainty in Neural Network for selecting the most informative sequences to query from labeling from a pool of unlabeled sequences. One is the most-possible-constraint-violation method (MPSV) and another is simple uncertainty (SU) method. We compare our proposed methods with random selection on three real data set from named entity recognition task and subphrase generation task for queries. For the task of entity recognition, our experiments reveal that this approach reduces annotation efforts in terms of manually labeled tokens compared to the standard, fully supervised AL scheme. Our experiment result on selecting sequences with same length shows that the most-possible-constraint-violation method (MPSV) and simple uncertainty (SU) outperform the random method significantly. Also MPSV outperforms SU by considering the underlying class distribution.

References

1. Lewis D., Gale W. A sequential algorithm for training text classifiers // Proc. of the ACM SIGIR Conf. on R & D in Information Retrieval. 1994. P. 3–12.
2. Settles B., Craven M., Ray S. Multiple-instance active learning // Advances in Neural Information Processing Systems (NIPS). 2008. Vol. 20. P. 1289–1296.
3. Seung H. S., Opper M, Sompolinsky H. Query by committee // Proc. of the ACM Workshop on Computational Learning Theory. 1992. P. 287–294.
4. Tong S., Koller D. Active Learning for Parameter Estimation in Bayesian Networks // NIPS. 2000. P. 647–653.
5. Cohn D., Ghahramani Z., Jordan M. I. Active learning with statistical models // J. of Artificial Intelligence Research. 1996. Vol. 4. P. 129–145.
6. Roy N., McCallum A. Toward optimal active learning through sampling estimation of error reduction // Proc. of the Intern. Conf. on Machine Learning (ICML). 2001. P. 441–448.
7. Settles B., Craven M. An analysis of active learning strategies for sequence labeling tasks // Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2008. P. 1069–1078.
8. Lafferty J., McCallum A., Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data // Proc. of the 18th Intern. Conf. on Machine Learning. 2001. P. 282–289.
9. Rabiner L. R. A tutorial on hidden markov models and selected applications in speech recognition // Proc. of the IEEE. 1989. Vol. 77. № 2. P. 257–286.
10. McCallum A., Freitag D., Pereira F. Maximum entropy Markov models for information extraction and segmentation // Proc. of the 17th Intern. Conf. on Machine Learning. 2000. P. 591–598.
11. Sang E. F. Introduction to the conll-2002 shared task: Language-independent named entity recognition // Proc. of the CoNLL-2002. 2002. P. 155–158.
12. Stegeman L. Part-of-speech tagging and chunk parsing of spoken Dutch using support vector machines // Proc. of the 4th Twente Student Conf. on IT. 2006.
13. Engel E. A. Modified artificial neural network for information processing with the selection of essential connections : Ph. D. thesis. Krasnoyarsk, 2004.
14. Tong S., Koller D. Support vector machine active learning with applications to text classification // Proc. of the ICML-00, 17th Intern. Conf. on Machine Learning. 2000. P. 999–1006.

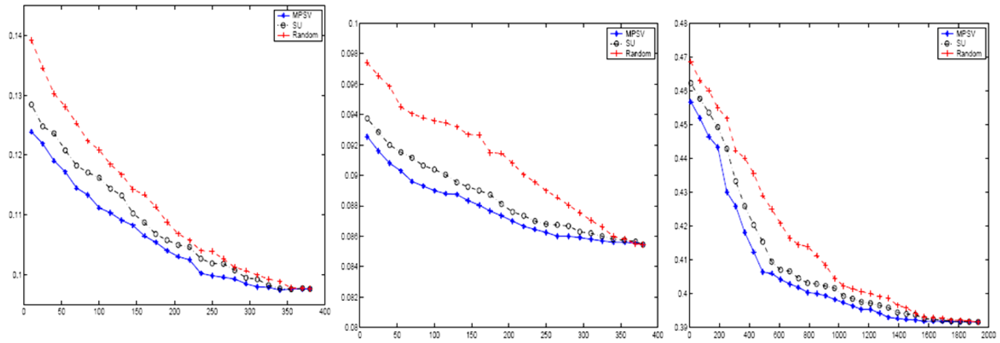

Fig. 2. The average error for ESP data set by three active learning uncertainty measurements