Query Selection via Weighted Entropy in Graph-Based Semi-Supervised Classification

Krikamol Muandet[1], Sanparith Marukatat[2], and Cholwich Nattee[1]

[1] School of Information, Computer and Communication Technology
Sirindhorn International Institute of Technology
Thammasat University, 131 Tiwanont Rd. Bangkadi
Pathum Thani, Thailand 12000

[2] National Electronic and Computer Technology Center
National Science and Technology Development Agency
111 Thailand Science Park
Pathum Thani, Thailand 12120

Abstract. There has recently been a large effort in using unlabeled data in conjunction with labeled data in machine learning. Semi-supervised learning and active learning are two well-known techniques that exploit the unlabeled data in the learning process. In this work, the active learning is used to query a label for an unlabeled data on top of a semi-supervised classifier. This work focuses on the query selection criterion. The proposed criterion selects the example for which the label change results in the largest pertubation of other examples' label. Experimental results show the effectiveness of the proposed query selection criterion in comparison to existing techniques.

Key words: Graph-Based Semi-supervised Learning, Active Learning

1 Introduction

There has recently been a large effort in using unlabeled data in conjunction with labeled data in machine learning. Indeed, in supervised learning, a huge amount of labeled data, e.g., manual annotation of legitimate and spam emails, is needed to train an accurate classifier. Labeled examples are however expensive and difficult to obtain since they require experienced annotators. Unlabeled instances, on the other hand, are easy to collect. Although they are usually useless in traditional supervised learning, unlabeled data may contain relevant information that can produce considerable improvement in learning accuracy. This is why several works [1–4] focus on how to exploit both labeled and unlabeled data to train the classifier. Semi-supervised learning and active learning are two well-known settings that have been considered to incorporate both kinds of data in learning process.

Semi-supervised learning has drawn large attention in the past decade due to its significance in many real-world problems. With few labeled examples, the

goal of this learning method is then to create as accurate as or even better classifier than what we would obtain from traditional supervised learning. Many successful techniques for semi-supervised learning have been proposed in various frameworks, e.g., generative models [5, 6], Gaussian processes [2, 7, 8], and information regularizations [9, 10]. Amongst these tecniques, semi-supervised learning on graph has received the most attention, as shown by the number of works on this subject including graph mincut [11, 12], learning with local and global consistency [13], and manifold regularizations [14, 15], for example. Interested readers should consult [16] for an extensive literature review of semi-supervised learning.

The goal of active learning coincides with that of semi-supervised learning, which aims at reducing the use of labeled data. The interesting aspect of active learning is that the learning system actively queries the instance whose label will be assigned by human annotators. As a result, the number of labeled examples required is usually less than what would actually be when learned using normal supervised learning. Active learning has been studied for many real-world problems such as text classification [17, 18]. Further information on active learning can be obtained from [19].

In this work, we propose the novel algorithm based on the combination of active learning and semi-supervised classification. More precisely, the system actively queries an instance from a pool of unlabeled instances. The human annotator will give the true label of this example. Then labeled examples along with the rest of unlabeled instances are used in standard semi-supervised classification. Due to its capability of evaluating the model successively, this learning framework becomes more effective than traditional semi-supervised learning processes. In this work we introduce a novel evaluative measure called *weighted entropy* that is used as a criterion for active query selection on top of graph-based semi-supervised classification.

This paper is organized as follows. Section 2 briefly reviews the related works. Next, we discuss the problem of graph-based semi-supervised learning in Sec. 3. The query selection algorithm based on the weighted entropy is then presented in Sec. 4. Experimental results are shown in Sec. 5, followed by conclusions in Sec. 6.

2 Related Works

All active learning techniques focus on how to find the optimal query, which generally involves evaluating the informativeness of unlabeled examples. There have been many proposed querying strategies that work well in practice. Random sampling [20] is the simplest querying strategy. This strategy is often used for preliminary testing of learning algorithms in active learning since it is very easy to implement. However, random method is not very effective in practice. For example, if the dataset contains much less number of positive examples than negative examples, then it is likely that a random sampling will select less positive examples than negative examples. As a result, Lewis and Gale [17] proposed

a querying framework called uncertainty sampling. In this framework, the algorithm asks an annotator to label those examples whose class membership is the most uncertain. This strategy significantly reduces the amount of training data that would have to be manually labeled to achieve a desired level of accuracy. However, uncertainty sampling is prone to query outliers, which may not be "representative" of other examples in the distribution [21, 22].

Due to this problem, the estimated error reduction framework has been proposed. It prefers the queries that minimize the expected future error of the learning system. Roy and McCallum applied this framework in text classification using Naive Bayes [21]. Zhu et al. [22] combined it with a semi-supervised classification. Both techniques result in an improvement over both random and uncertainty sampling. However, the estimated error reduction may be the most expensive query selection framework for many model classes since it requires estimating the expected future error over the set of unlabeled data for each query. This usually means that a new model must be incrementally trained for each possible query, leading to an increase in computational cost for some models, such as a logistic regression models and neural networks. Fortunately, this is not the case for graph-based semi-supervised classification with Gaussian random field models [22] for which the incremental training technique is exact and efficient. Therefore, combining active learning with semi-supervised classification guarantees that this approach will be fairly practical.

3 Graph-Based Semi-supervised Classification

Graph-based semi-supervised classification methods utilize a weighted graph whose nodes represent labeled and unlabeled data points. The weighted edges reflect the similarities between nodes. Most existing algorithms are based on the assumption that the labeling is smooth on the graph, i.e., similar data points tend to have similar labels. This is called the *smoothness assumption* [3].

Let $\mathcal{X} = \{x_1, \ldots, x_l, x_{l+1}, \ldots, x_n\} \subset \mathbb{R}^m$ be a set of n data points. The first l data points are labeled as $y = [y_1, y_2, \ldots, y_l]^{\mathrm{T}}$ with $y_i \in Y = \{0, 1\}$. The rest of $u = n - l$ data points are initially unlabeled. Let L and U denote sets of labeled and unlabeled examples, respectively. The goal of semi-supervised learning is to utilize labeled data together with unlabeled data to construct a classifier.

The graph is represented by a matrix $W = [w_{ij}]_{n \times n}$. The non-negative edge weight w_{ij} between node i, j is computed as

$$w_{ij} = \exp(-\sum_{d=1}^{m}(x_{i,d} - x_{j,d})^2/\sigma_d^2) \ , \tag{1}$$

where $x_{i,d}$ is the d-th component of x_i and σ_d is the bandwidth hyperparameters for the dimension d.

In general, graph-based semi-supervised classification technique searches for a real-valued function f on graph and then assigns labels based on f. Given the weight matrix W and the real-valued function f, the inconsistency on graph can be define according to the smoothness assumption as

$$E(f) = \sum_{i,j=1}^{n} w_{ij}(f(i) - f(j))^2 \ , \tag{2}$$

where $f(i)$ and $f(j)$ are the function values evaluated on node i and j, respectively. For a labeled example $1 \leq i \leq l$, $f(i)$ is considered fixed to y_i. The inconsistency term (2) can also be written in quadratic form $f^{\mathrm{T}}\Delta f$, where $f = [f(1), .., f(n)]^{\mathrm{T}}$ and Δ is known as the combinatorial graph Laplacian matrix defined as $\Delta = D - W$, where the matrix D is a diagonal matrix whose entries $D_{ii} = \sum_{j=1}^{n} w_{ij}$ is the degree of node i. Minimization of $E(f)$ forces f to take values y_i on labeled data points and varies smoothly on unlabeled data points in accordance with the weight matrix W. It is not difficult to show that the optimal function is given by:

$$f_U = (D_{UU} - W_{UU})^{-1} W_{UL} f_L = -\Delta_{UU}^{-1} \Delta_{UL} f_L \ , \tag{3}$$

where f_U is a vector of function values evaluated on all unlabeled data points, $f_L = [y_1, y_2, \ldots, y_l]^{\mathrm{T}}$, and $f = [f_L; f_U]$. All related matrices are defined as follows:

$$W = \begin{pmatrix} W_{LL} \ W_{LU} \\ W_{UL} \ W_{UU} \end{pmatrix}, \Delta = \begin{pmatrix} \Delta_{LL} \ \Delta_{LU} \\ \Delta_{UL} \ \Delta_{UU} \end{pmatrix} \ .$$

The functional (3) is called the *soft-label* function since its values do not directly specify the class membership of unlabeled examples, but can be interpreted as a probability of being in each class. The most obvious method to transform soft-label to hard-label is by thresholding, e.g., classify x_i, $l \leq i \leq n$, as being in class 1 if $f(i) > 0.5$, and in class 0 otherwise. This method generally works well when the classes are well-separated. However, this is generally not the case in many practical applications. In such cases, using simple thresholding may result in an unbalanced classification.

Class Mass Normalization (CMN) is another method to transform the soft-label to hard-label [3]. The class distribution of the data is adjusted to match the class priors, that can be obtained from the labeled examples. For example, if the prior class proportion of class 1 and 0 is p and $1 - p$, respectively, then an unlabeled examples x_k is classified as class 1 iff $p \cdot (f(k)/\sum_i f(i)) > (1 - p) \cdot ((1 - f(k))/\sum_i (1 - f(i)))$. This method works well when we have sufficient labeled examples to determine the class prior that accurately represents the true class distribution.

4 Weighted Entropy

We propose a novel technique to perform active learning with graph-based semi-supervised learning. This active learning technique selects queries from the unlabeled data by considering their influence on other available examples. Each unlabeled example is evaluated by looking at the characteristics of the overall

soft-label function when its label is altered. That is, we prefer the example that if its label is changed, will result in (1) the large change in soft-label values of other unlabeled examples that can probably alter their class membership and (2) the large change in soft-label values of other unlabeled examples that rarely changes the class membership of other examples, but increases the confidence of the labeling function. We propose *weighted entropy* as an evaluation function which can simultaneously take both criteria into account. The criterion function is derived from the harmonic energy minimization function with the Gaussian random field model [3].

4.1 Problem Formulation

In graph-based setting, we need to know a soft-label function when different labels are assigned to unlabeled node k to evaluate its impact to other unlabeled examples. Thus it is necessary to define an efficient update of soft-label function after knowing one more label. Following the derivation in [22], we add one new node with value f_0 to the graph. The new node is connected to unlabeled node k with weight w_0. Thus, as $w_0 \to \infty$, we effectively assign label f_0 to node k.

Note that the harmonic energy minimization function is $f_U = -\Delta_{UU}^{-1}\Delta_{UL}f_L = (D_{UU}-W_{UU})^{-1}W_{UL}f_L$. After adding the new node to the graph, let the matrices $\tilde{D}_{UU}$,$\tilde{W}_{UL}$, and $\tilde{f}_L$ be the updated versions of D_{UU}, W_{UL}, and f_L, respectively. Since the new node is a labeled node in the graph, the labeling function can be updated as

$$\begin{aligned}\tilde{f}_U &= (\tilde{D}_{UU} - W_{UU})^{-1}\tilde{W}_{UL}\tilde{f}_L \\ &= (w_0 e_k e_k^{\mathrm{T}} + D_{UU} - W_{UU})^{-1}(w_0 f_0 e_k + W_{UL}f_L) \\ &= (w_0 e_k e_k^{\mathrm{T}} + \Delta_{UU})^{-1}(w_0 f_0 e_k + W_{UL}f_L)\ , \end{aligned} \tag{4}$$

where e_k is a column vector with 1 in position k and 0 elsewhere. Using matrix inversion lemma, we obtain

$$\begin{aligned}(w_0 e_k e_k^{\mathrm{T}} + \Delta_{UU})^{-1} &= \Delta_{UU}^{-1} - \frac{\Delta_{UU}^{-1}(\sqrt{w_0}e_k)(\sqrt{w_0}e_k)^{\mathrm{T}}\Delta_{UU}^{-1}}{1 + (\sqrt{w_0}e_k)^{\mathrm{T}}\Delta_{UU}^{-1}(\sqrt{w_0}e_k)} \\ &= \mathcal{L} - \frac{w_0\mathcal{L}_{|k}\mathcal{L}}{1 + w_0\mathcal{L}_{kk}}\end{aligned} \tag{5}$$

where we use $\mathcal{L}$ to denote Δ_{UU}^{-1} and $\mathcal{L}_{|k}$ is a square matrix with $\mathcal{L}$'s k-th column and 0 elsewhere. With some calculations, we derive (5) into

$$\tilde{f}_U = f_U + \frac{w_0 f_0 - w_0 f(k)}{1 + w_0\mathcal{L}_{kk}}\mathcal{L}_{\cdot k} \tag{6}$$

where $f(k)$ is the soft-label of unlabeled node k, and $\mathcal{L}_{\cdot k}$ is the k-th column vector in $\mathcal{L}$. To force the label at node k to be f_0, we let $w_0 \to \infty$ to obtain

$$\tilde{f}_U = f_U + \frac{f_0 - f(k)}{\mathcal{L}_{kk}} \mathcal{L}_{\cdot k} \ . \tag{7}$$

Consequently, we can now formulate the equation to compute the soft-label function after knowing the label of a particular example. The functional (7) can be used to compute the possible soft-label functions after assigning different labels to example x_k. For binary classification in which $Y = \{0, 1\}$, we obtain

$$f_U^{(x_k,y)} = f_U + (y - f(k)) \frac{(\Delta_{UU}^{-1})_{\cdot k}}{(\Delta_{UU}^{-1})_{kk}} \ , \tag{8}$$

where $f_U^{(x_k,y)}$ is the soft-label function after we assign label $y \in Y$ to unlabeled examples x_k.

To derive the evaluation function, we start by defining the uncertainty $I(f_{U,i})$ of the soft-label value at node i. The total uncertainty at node i when we assign different labels to example x_k is $\sum_{y \in Y} I(f_{U,i}^{(x_k,y)})$. This work uses entropy function to measure this uncertainty, i.e., $I(f_{U,i}) = -f_{U,i} \log_2 f_{U,i}$. By weighting each uncertianty term with the probability of x_k belonging to each class, we can write the evaluation function as

$$\begin{aligned} \mathcal{C}(k) &= \sum_{i=1}^{u} \sum_{y=0,1} p(y_k = y|L) I(f_{U,i}^{(x_k,y)}) \\ &= \sum_{i=1}^{u} \sum_{y=0,1} -p(y_k = y|L) \left[f_{U,i}^{(x_k,y)} \log_2 f_{U,i}^{(x_k,y)} \right] \ . \end{aligned} \tag{9}$$

This quantity is called the *weighted entropy* measure. Given a set of label data L, $p(y_k = y|L)$ is the true label distribution at example x_k. This term makes $\mathcal{C}(k)$ not computable, however it can still be estimated by the soft-label value, i.e. , $p(y_k = 1|L) \approx f_{U,k}$ and $p(y_k = 0|L) \approx 1 - f_{U,k}$. As a result, the *estimated weighted entropy* is defined as

$$\hat{\mathcal{C}}(k) = \sum_{i=1}^{u} -f_{U,k} \left[f_{U,i}^{(x_k,1)} \log_2 f_{U,i}^{(x_k,1)} \right] - (1 - f_{U,k}) \left[f_{U,i}^{(x_k,0)} \log_2 f_{U,i}^{(x_k,0)} \right] \ . \tag{10}$$

The weighted entropy measures how much the selected query affects the labels of other unlabeled examples when its own label is changed. Note that the soft-label function determines the probability of examples being in positive class, i.e., $y_i = 0$ if $f(i) \leq 0.5$ and $y_i = 1$ otherwise.

4.2 Analysis of Weighted Entropy

In this work, the query is selected in accordance with the influence when its label is changed. Given a particular example x_k, the behavior of weighted entropy can be analyzed as follows:

1. The value of weighted entropy is minimized when the expected soft-label values $f_U^{(x_k,0)}$ and $f_U^{(x_k,1)}$ have large difference, e.g., $f_U^{(x_k,1)}$ is close to 1 and $f_U^{(x_k,0)}$ is close to 0, or vice versa. This means that the example x_k has a significant effect on the soft-label values of other unlabeled examples. Furthermore, such effect will likely cause the alternation of class membership of most examples. As a result, this suggests that we should know the true label of x_k as early as possible.
2. The value of weighted entropy is also minimized when both $f_U^{(x_k,1)}$ and $f_U^{(x_k,0)}$ become closer to the boundary value of the soft-label function. This means that even if changing the label of x_k does not alter labels of other examples but if it results in more confident labeling function, then x_k will be considered as an important example by weighted entropy measure. In this case, we can see that the formulation of estimated weighted entropy (10) is similar to the expected estimated risk defined in [22]. The expected estimated risk is computed by summing a weighted estimated risks of $f_U^{(x_k,y)}$ using $\min(f_U^{(x_k,y)}, 1-f_U^{(x_k,y)})$. In contrast, by using the weighted entropy, the proposed method focuses on the behavior of soft-label values of unlabeled data when the label of queried example is changed. Therefore, this method can effectively exploit necessary information needed to evaluated unlabeled data.

It is worth mentioning that we expect these two cases to happen in different periods during the query selection process. It is obvious that the queries corresponding to the first case will likely become the most preferable candidates at the beginning of the process, when labels are prone to change. After some times when labels become more resistant to change, the most preferable candidates will fall into the second case, which tends to improve the confidence of the classifier. Therefore, selecting the queries based on the proposed criterion assures that examples that have negative effects on the performance as a result of learning process will be discovered as early as possible. Consequently, the query selection will become safer for the subsequent learning process. The final query selection criterion (11) called *Minimum Weighted Entropy* (MinWE) for query selection is shown below.

Minimum Weighted Entropy
For a set of unlabeled examples U, select an example x_k resulting in

$$k = \arg\min_{k'} \ \hat{\mathcal{C}}(k') \ , \tag{11}$$

where $\hat{\mathcal{C}}(k')$ is the estimated weighted entropy of $x_{k'}$.

Algorithm 1 summarizes the active query selection using minimum weighted entropy. Note that in each iteration we need to update the inverse graph Laplacian with the row/column for x_k removed, whose computational complexity is $\mathcal{O}(u^3)$ in general. To avoid the computational cost of the matrix inversion, the

Algorithm 1 Active Query Selection with Minimum Weighted Entropy

1: Choose first r examples randomly.
2: **while** Need more queries **do**
3: Update the inverse graph Laplacian Δ_{UU}^{-1}.
4: Compute $\hat{\mathcal{C}}(k)$ using (10) for all $x_k \in U$.
5: Query x_k according to (11).
6: Receive the answer y_k.
7: Add (x_k, y_k) into L and remove x_k from U.
8: **end while**

matrix inversion lemma is applied to compute this matrix from Δ_{UU} and Δ_{UU}^{-1} (see appendix B of [22] for the derivation). As a result, the overall time complexity of the proposed algorithm is $\mathcal{O}(u^2)$.

5 Experimental Results

5.1 Experimental Setup

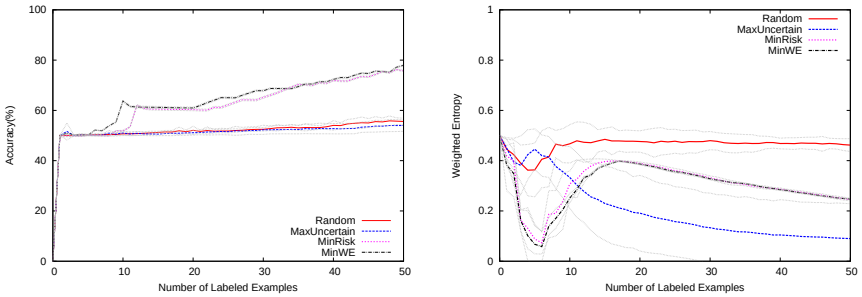
In our experiments, `Random`, `MaxUncertain`, `MinRisk`, and `MinWE` query selection methods are evaluated on different datasets, namely, handwritten digits dataset[3] and benchmark datasets for semi-supervised learning[4]. All experiments are performed in the transductive setting, i.e., test data coincides with training data used in the learning process. In each experiment, the first two labeled examples are chosen by random selection. Since random initialization can consequentially influence the results, each experiment is repeated 10 times and the average accuracy as well as its ± 1 standard deviation are reported.

In `Random` method, the next instance is queried randomly from a set of available unlabeled instances. In contrast, the `MaxUncertain` method queries an unlabeled instance whose soft-label value is closest to 0.5, meaning that the instance possesses the highest uncertainty on its true label. The `MinRisk` method selects an instance that minimizes the expected estimated risk defined in [22]. Similarly, the `MinWE` method queries an instance that minimizes the estimated weighted entropy proposed in this work. For comparison, the classification accuracy attained by each method is evaluated on the unlabeled instances at each iteration after adding the new query to the set of labeled examples.

5.2 Overview of Results

Figure 1 shows the accuracy and weighted entropy values on handwritten digits dataset as increasing number of labeled examples are acquired using different query selection methods. The task is to classify the digit "0" against the digit "8". The digits are 16×16 grid, with pixel values ranging from 0 to 255. Thus

[3] http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html
[4] http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html



(a) The accuracy (b) The weighted entropy value

Fig. 1. (a) Accuracy and (b) the weighted entropy values on the handwritten digits dataset("0" versus "8" classification problem) of four different query selection methods.

each digit is represented by a 256-dimensional vector. The dataset consists of 1,080 images, with 540 images randomly selected from each class. All data points and their pairwise similarities are represented by a fully connected graph with weights $w_{ij} = \exp(-d_{ij}^2/0.25)$, where d_{ij} is the Euclidean distance between x_i and x_j.

As illustrated in Fig.1(a), `Random` and `MaxUncertain` methods achieve approximately the same level of accuracy, which increases very slowly as the system acquires more labeled examples. In contrast, the accuracy attained by `MinRisk` and `MinWE` methods rises very rapidly. Although they lead to roughly the same results in the later stage of the learning process, the `MinWE` method discloses the informative queries earlier than the `MinRisk` method as depicted at the 10^{th} query in Fig.1(a).

Fig.1(b) shows the weighted entropy value versus the number of labeled examples. The `MinRisk` and `MinWE` methods generate the similar trend of weighted entropy values, which tend to decrease with the increasing number of labeled examples. Though exhibiting the similar criteria, the `MinWE` method produces a slighly lower values of weighted entropy than the `MinRisk` method. Another interesting point in this experiment is the weighted entropy values of `MaxUncertain` method that decrease substantially. According to Fig.3, this case usually occurs when the values of soft-label are very close to either 0 or 1 because the majority of queries come from the same class. This degenerate case of `MaxUncertain` method primarily leads to low accuracy.

Note that an advantage of `MinWE` method over `MinRisk` method is the ability to initially identify important queries that have significant impact on the soft-label value of other unlabeled examples if their labels are changed. Thus it follows immediately that imposing the true labels on these examples lessens the label alteration in the following stages. The `MinRisk` method choose queries based on how much they will improve the expected accuracy. It does not take into account the fact that this improvement can change some of influential examples' labels that probably degrades the true accuracy.

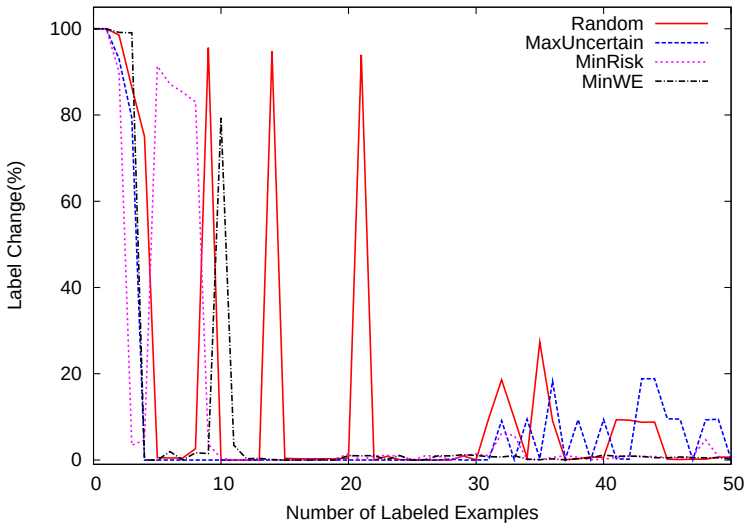


Fig. 2. The percentage of labels on unlabeled examples that change as the increasing number of labeled examples is obtained in handwritten digits dataset

Figure 2 supports our previous claim on the advantages of `MinWE` method. The percentage of label changes in a set of unlabeled examples of handwritten digits dataset is calculated at each iteration after the next query is obtained. Therefore, this figure nearly reflects the reliability of each query selection method, i.e., the reliable method rarely alters the labels of examples compared to unreliable ones. Intuitively, the `Random` method is unreliable because the labels frequently change throughout the learning process. Nonetheless, the `MinRisk` and `MinWE` methods are more reliable since the variation of labels occurs essentially in the initial stages of the learning process and then becomes less likely to occur in the subsequent stages. Moreover, although `MinRisk` method substantially lessens the generalization error and the the alteration of labels in the beginning, the correctness of labeling function at some data points may not be guaranteed. As a result, obtaining more label examples likely changes labels of some examples as illustrated in Fig. 2 when there are approximately 8, 32, and 48 labeled examples. Note that, on the other hand, the labels of most influential queries are assigned in the initial stages by `MinWE` method. Therefore, the variation of labels in the subsequent stages is minimized, leading to more reliable labeling function.

Another issue we want to address in this work is the meaningfulness of the queries selected by each methods. It is worthwhile to mention because it also affects the performance of the classifier. If the selected queries are ambiguous, it is with high chance that the assigned labels will be incorrect and knowing their labels will not provide any useful information. This problem may not be recognized in easy tasks such as handwritten recognition or face detection, for

Fig. 3. Top twenty most frequent queries of the handwritten digits obtained during the learning process using `Random`, `MaxUncertain`, `MinRisk`, and `MinWE` methods.

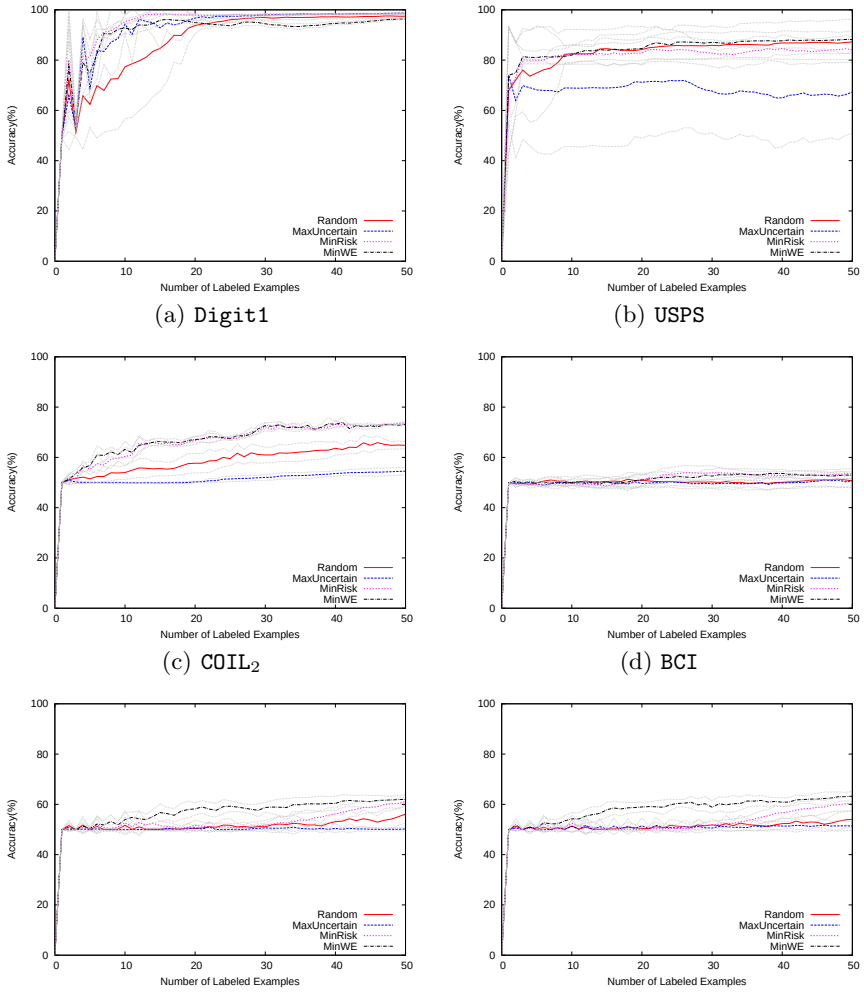
example, but it becomes more realistic when we handle the complicated problems, in which it is not convenient to visualize the data. Figure 3 illustrates the 20 most frequently queried instances across all trials. Each row shows the most frequent queries, sorted by the number of times they are selected.

Figure 4 shows the results of the experiments on benchmark datasets for semi-supervised learning. The original benchmark consists of eight datasets, but in this work we use only six of them to assess the query selection methods[5]. The datasets are categorized into two groups. The first group consists of `g241c`, `g241n`, and `Digit1`, which are artificially created. The second group includes `USPS`, $\texttt{COIL}_2$, and `BCI`, all of which are derived from real data. In addition, the classes of `USPS` dataset are imbalanced with relative sizes of 1:4. Thus the performance on these datasets is the indication of the performance in the real applications. See [23] for the detail of each dataset. In this experiment, we construct a weighted k nearest neighbors graph with weights $w_{ij} = \exp(-d_{ij}^2/2\sigma)$ if it is an edge between x_i and x_j, and 0 otherwise. For all datasets, $k = 5$ and σ is fixed as the median of the pairwise distance between adjacent nodes on the graph.

Figure 4 confirms an effectiveness of the proposed query selection method. As can be seen in figure, the `MinWE` method outperforms other query selection methods and is slighly better than `MinRisk` method in almost all datasets, except `Digit1` dataset. Other than the superior experimental results on artificial datasets such as `g241c` and `g241n`, `MinWE` method also achieve the highest accuracy in `USPS`, $\texttt{COIL}_2$, and `BCI` datasets that are obtained from the real data. This is therefore the indication of expected classification performance of `MinWE` method in practice.

In almost all datasets, the `MaxUncertain` query selection method possesses the lowest accuracy among all other methods. In some datasets, e.g., `USPS`, labeling the most uncertain unlabeled example may harm the accuracy as illustrated

[5] The `g241c` , `g241n`, `Digit1`, `USPS`, $\texttt{COIL}_2$, and `BCI` datasets are used extensively to assess the performance of several semi-supervised learning techniques, whereas few techniques utilize `Text` and `SecStr` datasets, which have special characteristics. Therefore, this work considers only those six datasets for the convenience of the experiments and the reliability of the results.



(a) `Digit1` (b) `USPS`

(c) $\texttt{COIL}_2$ (d) `BCI`

(e) `g241c` (f) `g241n`

Fig. 4. The accuracy of different query selection methods on benchmark dataset for semi-supervised learning. The selected datasets includes (a) `Digit1`, (b) `USPS`, (c) $\texttt{COIL}_2$, (d) `BCI`, (e) `g241c`, and (f) `g241n`. The figure presents an average accuracy and its ± 1 standard deviation.

in Fig. 4(b). This underlines the fact that in semi-supervised learning, more labeled examples does not mean higher accuracy. Intuitively, the uncertainty on class membership of a particular instance does not adequately indicate the informativeness of the instance. The evaluation based on this measure suggests only that the instance is hard to classify, but does not imply how, after knowing its label, the instance will influence the rest of unlabeled instances. Consequently,

Table 1. Statistical comparison of query selection methods. The average accuracy of query selection methods on each dataset (left) and corresponding results of Friedman and Wilcoxon signed ranks tests (right) are reported.

	Random	MaxUncertain	MinRisk	MinWE
Digit1	88.49	93.04	93.57	92.72
USPS	83.71	68.31	81.48	86.23
COIL2	59.06	51.68	65.96	70.44
BCI	50.34	49.91	51.72	52.71
g241c	51.67	50.22	52.30	59.48
g241n	51.41	50.92	52.01	59.38

	Test Statistics			
Friedman Test	N	Chi-Square	df	Sig.
	6	11.600	3	**.009**
Wilcoxon Test		Z		Sig.
MinWE-Random		-2.201		**.028**
MinWE-MaxUncertain		-1.992		**.046**
MinWE-MinRisk		-1.992		**.046**

in practice, we recommend to use `MaxUncertain` only for preliminary testing, especially in the difficult problems, for the following reasons:

1. The uncertainty on class membership does not suggest, neither directly nor indirectly, the influence of the queried instance on the labels of other instances. Thus providing its label may not considerably improve, but probably diminish the predictive accuracy.
2. The most uncertain instance is hard not only for the system, but also for the human annotators to categorize as illustrated in Fig.3. Hence, there is a high chance that the given label is incorrect or useless.

Fortunately, the `MinWE` query selection method does not suffer from these two problems as shown by the experimental results. That is, it evaluates the informativeness of those instances by the effect they make on the whole dataset. Therefore, labeling more instances guarantee to improve the predictive performance of the classification. It is also worth to note that the `MinWE` achieve better classification results than the `MinRisk` method even though its evaluative measure does not directly take into account the accuracy of classification, compared to the estimated risk defined in the `MinRisk`.

Statistical Comparison To justify that the proposed method significantly yields an improve performance over the existing query selection methods, we need to perform a proper statistical test over all datasets with the null hypothesis that all methods perform equally well. As suggested by [24] we used the Friedman and Wilcoxon signed ranks tests.

Friedman test is a nonparametric test used to compare three or more observations repeated on the same subjects. In this case, we compare the average learning accuracy of four query selection methods over six datasets to inspect the difference in medians between different methods. Under the null hypothesis, which states that all methods are equivalent, the Friedman test is found to be significant $\chi^2(3, N = 6) = 11.6$ and $p < .01$, as shown in Table 1. This merely indicates the differences in learning accuracy among the four methods.

Next, after obtaining a significant Friedman test, the follow-up tests need to be conducted to evaluate comparisons between pairs of query selection methods. As indicated earlier, we use the Wilcoxon signed ranks test. Since we are only

interested in whether the proposed method yields an improved performance over the existing ones, the pairwise comparisons between the `MinWE` method and the others, namely, `Random`, `MaxUncertain`, and `MinRisk` are conducted. The null hypothesis for each comparison states that there is no difference in learning accuracy between two methods, whereas alternative hypothesis states that the first method, i.e., `MinRisk`, gives higher accuracy than the second one. According to the result of Wilcoxon test in Table 1, all three comparisons are significant at the .05 alpha level, leading us to reject the null hypothesis and conclude that the `MinWE` method significantly outperforms `Random`, `MaxUncertain`, and `MinRisk` methods.

6 Conclusions

This paper proposes a new query selection criterion for active learning. The proposed criterion, called minimum weighted entropy, selects the example for which the label change results in the largest pertubation of other examples' label. It relies on a graph-based semi-supervised learning technique to efficiently compute the weighted entropy for the selection process. Experimental results show the advantage of the proposed selection criterion over the existing criterion on several datasets.

Acknowledgments This work is supported in part by the Young Scientist and Technologist Programme or YSTP (SIIT-NSTDA:S1Y48/F-002) of the National Science and Technology Development Agency, Thailand.

References

1. Krishnapuram, B., Williams, D., Xue, Y., Hartemink, A.J., Carin, L., Figueiredo, M.A.T.: On semi-supervised classification. In: NIPS. (2004)
2. Zhu, X., Ghahramani, Z.: Semi-supervised learning: From gaussian fields to gaussian processes. Technical report, School of CS, CMU (2003)
3. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: In ICML. (2003) 912–919
4. Minton, S., Knoblock, C.A.: Active + semi-supervised learning = robust multi-view learning. In: Proceedings of ICML-02, 19th International Conference on Machine Learning. (2002) 435–442
5. Nigam, K., Mccallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. In: Machine Learning. (1999) 103–134
6. Baluja, S.: Probabilistic modeling for face orientation discrimination: learning from labeled and unlabeled data. In: Proceedings of the 1998 conference on Advances in neural information processing systems II, Cambridge, MA, USA, MIT Press (1999) 854–860
7. Lawrence, N.D., Jordan, M.I.: Semi-supervised learning via gaussian processes. In Saul, L.K., Weiss, Y., Bottou, L., eds.: Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA (2005) 753–760

8. Chu, W., Sindhwani, V., Ghahramani, Z., Keerthi, S.S.: Relational learning with gaussian processes. In Schölkopf, B., Platt, J., Hoffman, T., eds.: Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007) 289–296
9. Szummer, M., Jaakkola, T.: Information regularization with partially labeled data. In: In Advances in Neural Information Processing Systems 15, MIT Press (2003) 2003
10. Tommi, A.C., Jaakkola, T.: On information regularization. In: In Proceedings of the 19th UAI, UAI (2003)
11. Blum, A., Lafferty, J., Rwebangira, M.R., Reddy, R.: Semi-supervised learning using randomized mincuts. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning, New York, NY, USA, ACM (2004) 13
12. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 19–26
13. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems 16, MIT Press (2003) 321–328
14. Belkin, M., Matveeva, I., Niyogi, P.: Regularization and semi-supervised learning on large graphs. In: In COLT, Springer (2004) 624–638
15. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research **7** (2006) 2399–2434
16. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison (2005) http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.
17. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, Springer-Verlag New York, Inc. (1994) 3–12
18. Mccallum, A.K.: Employing em in pool-based active learning for text classification. In: In Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann (1998) 350–358
19. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
20. Cochran, W.G.: Sampling Techniques. John Wiley and Sons, New York (1977)
21. Roy, N., Mccallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: In Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann (2001) 441–448
22. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. (2003) 58–65
23. Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
24. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7** (2006) 1–30