Graph-based Iterative Hybrid Feature Selection

ErHeng Zhong[†] Sihong Xie[†] Wei Fan[‡] Jiangtao Ren[†] * Jing Peng[§] Kun Zhang[$]

[†]Sun Yat-Sen University, {sw04zheh, mc04xsh, issrjt}@mail2.sysu.edu.cn

[‡]IBM T. J. Watson Research Center, weifan@us.ibm.com

[§]Montclair State University, pengj@mail.montclaire.edu

[$]Xavier University of Louisiana, kzhang@xula.edu

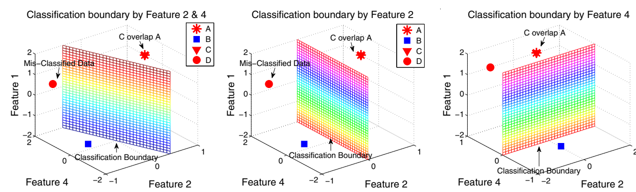
Abstract

When the number of labeled examples is limited, traditional supervised feature selection techniques often fail due to sample selection bias or unrepresentative sample problem. To solve this, semi-supervised feature selection techniques exploit the statistical information of both labeled and unlabeled examples in the same time. However, the results of semi-supervised feature selection can be at times unsatisfactory, and the culprit is on how to effectively use the unlabeled data. Quite different from both supervised and semi-supervised feature selection, we propose a "hybrid" framework based on graph models. We first apply supervised methods to select a small set of most critical features from the labeled data. Importantly, these initial features might otherwise be missed when selection is performed on the labeled and unlabeled examples simultaneously. Next, this initial feature set is expanded and corrected with the use of unlabeled data. We formally analyze why the expected performance of the hybrid framework is better than both supervised and semi-supervised feature selection. Experimental results demonstrate that the proposed method outperforms both traditional supervised and state-of-the-art semi-supervised feature selection algorithms by at least 10% in accuracy on a number of text and biomedical problems with thousands of features to choose from. Software and dataset is available from the authors.

1 Introduction

Traditional supervised feature selection methods can fail when the number of labeled examples is limited due to sample selection bias [12]. Semi-supervised methods attempt to correct this problem by exploiting unlabeled examples.

* The author is supported by the National Natural Science Foundation of China under Grant No. 60703110



(a) (b) (c)

Figure 1. Illustration of Toy Example

However, the success of these techniques depends mainly on a distance measure that provides an estimate of similarity between examples. Thus, a poor distance measure can seriously distort information obtained from the unlabeled data, thereby making the semi-supervised algorithms unreliable. To illustrate the problem of using either supervised or semi-supervised feature selection alone, we use the following example. Assume that there are four instances A, B, C and D. Each has four features: A(1,1,1,1), B(1,-1,1,-1), C(-1,1,1,1), and D(0,-1,1,1). Further, assume that A has label -, B has label +. Both C and D have a true label -, but for the moment, their labels are withheld from the feature selection process. As shown in Fig 1(c), one can observe that only feature 4 is exactly correlated to the class label, or the true label is positive if feature 4 has value -1. An ideal feature selection algorithm ought to choose feature 4 from all features. A supervised feature selection method using only A and B will select features 2 and 4. Based on labeled data, both feature 2 and feature 4 can differentiate data points from different classes. Supervised feature selection cannot rank one feature higher than the other, so both are selected. However, only feature 4 is actually useful. As shown in Fig 1(a), D is classified incorrectly using features 2 and 4. Now consider semi-supervised feature selection, if we use both the labeled data, A and B, and the unlabeled data, C and D, to select features simultaneously in the whole feature space, feature 2 will be selected if only one feature is desired. When we classify C and D using the classification boundary built from this selected feature, D will be on the same side as C, which is incorrect (Fig 1(b)). The moral of the story is that using either supervised or semi-supervised methods alone may miss critical features (i.e. feature 4). Next we propose a hybrid approach to solving these problems. We take advantage of the strengths of both the supervised and semi-supervised feature selection paradigms, while addressing their deficiencies. The key process is summarized in Fig 2. We first perform supervised feature selection on the labeled data to obtain an initial "seed" feature subset. This feature subset is then used to construct a more effective distance measure to separate the unlabeled data, as we shall see in Section 3. During each iteration, we improve this feature subset using the unlabeled data. To be exact, we use a graph model encoding the relationship between instances to predict the unlabeled data [14]. A new training set is built that includes both the labeled examples and those unlabeled examples whose predicted labels are likely to be correct. Feature selection will be performed on the new training set, and selected features will be used to construct a new graph model for the next iteration. We provide a formal analysis (Section 3) to justify the supervised feature selection before using unlabeled data to either reinforce or discard the selected features. To illustrate the advantage of the hybrid framework, we re-visit the example in Fig 1(a). First, we perform supervised feature selection on A and B, resulting in features 2 and 4 being selected. After that we use a graph model to label C and D using features 2 and 4, resulting in predicted class label -. We then perform semi-supervised feature selection on A, B, C and D. We find that only feature 4 is useful and feature 2 can be removed. This is because after C and D are labeled as -, only feature 4 can correctly separate these 4 examples while feature 2 can't. This complete process is illustrated Fig 1(c).

In summary, the hybrid feature selection framework has the following main advantage over both supervised and semi-supervised feature selections: (1) It performs supervised feature selection before predicting on unlabeled data, thereby maintaining the most critical features and providing better confidence estimates (as demonstrated in Section 3). Thus, the unlabeled data are used more selectively than semi-supervised approaches. (2) It is flexible, and can incorporate many feature selection methods that aim at removing irrelevant and redundant features and revealing the relationship among data points.

2 Hybrid Graph Model Feature Selection

We first give a short introduction to a graph-based classification algorithm [14], and then present details on the IteraGraph_FS (Iterative Hybrid Graph-based Semi-Supervised Feature Selection) algorithm.

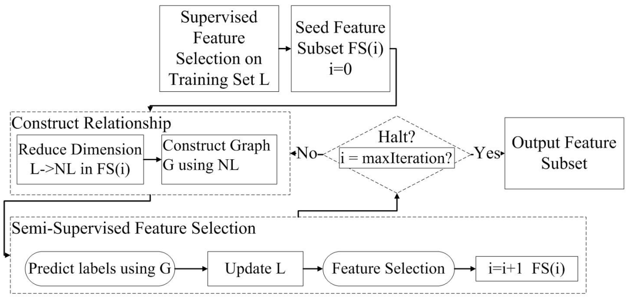


Figure 2. Illustration of our framework

Label Propagation We adopt the label propagation scheme for the labeled and unlabeled data proposed in [14]. Assume there are ℓ labeled points $(x_1, y_1), \ldots, (x_\ell, y_\ell)$, and u unlabeled points $x_{\ell+1}, \ldots, x_{\ell+u}$. For simplicity, we focus on two class problems, i.e., $y \in 0, 1$. We construct a connected graph $G = \{V, E\}$ with nodes V corresponding to the labeled and unlabeled data examples. Specifically, nodes $L = \{1, \ldots, \ell\}$ represent the labeled points with labels $y_1, \ldots, y_\ell$, while nodes $U = \{\ell + 1, \ldots, \ell + u\}$ represent the unlabeled points. The edges $(i, j) \in E$, $i, j = 1, \cdots, \ell + u$, are weighted according to

$$w_{ij} = w_{ji} = \exp(-\frac{\|x_i - x_j\|^2}{\lambda^2}) \tag{1}$$

where λ is a bandwidth hyper-parameter. It is adaptable according to the data. Following the analysis in [14], we set $\lambda = d^0/3$, where d^0 is the minimal distance between class regions. We define P to be the probabilistic transition matrix partitioned according to the labeled and unlabeled index L and U, $P = \begin{pmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{pmatrix}$ with entries $p_{ij} = p(i \rightarrow j) = w_{ij}/d_i$ where $d_i = \sum_{k \in N(i)} w_{ik}$ and $N(i)$ denotes the set of the neighbors of node i in the graph. Thus p_{ij} can be interpreted as the probability of making a transition from node i to node j. Similar to the ideas described in [14], we define a real-valued harmonic function $f : V \rightarrow \mathbb{R}$,

$$f(i) = \sum_{j \in N(i)} p_{ij} f(j), i = \ell + 1, \ldots, \ell + u \tag{2}$$

As suggested in [14], the label propagation algorithm in fact computes the harmonic function defined above. That is, $f = Pf$. Let $f = [f_l, f_u]^{\mathrm{T}}$ where f_u denotes the values on the unlabeled data and f_l represents the values on the labeled data (equal to the true label). Then the harmonic solution is

$$f_u = (I - P_{uu})^{-1} P_{ul} f_l \tag{3}$$

To understand Eq. (3), notice that since $\|P_{uu}\|_\infty < 1$, $(I - P_{uu})^{-1} = \sum_{i=0}^{\infty} P_{uu}^i$ (Neumann Series). Thus, $f_u = P_{ul} f_l + P_{uu} P_{ul} f_l + P_{uu}^2 P_{ul} f_l + P_{uu}^3 P_{ul} f_l + \cdots$

Algorithm 1 IteraGraph_FS

Input: L,U,$sizeFS$,s,$Iterations$
Output: $ResultFS$
FS=feature_selection(L) where $|FS| = sizeFS$
$newL = L$
for $i = 1$ **to** $Iterations$ **do**
  $G = creating(L + U, FS)$ using Eq. (3)
  $LabelsU$=G_predict(U,FS) through "Label Propagation" algorithm
  PU=top $s\%$ from $LabelsU$ with highest confidence
  $newL = L + PU$
  $newFS(i)$=feature_selection($newL$),where $|newFS| = sizeFS$
  $AvgConf(i)$=average prediction confidence on PU
end for
ResultFS=newFS(k) with the highest prediction confidence $AvdConf$

Iterative Hybrid Graph-based Semi-Supervised Feature Selection First, we perform supervised feature selection on the labeled data L, and obtain an initial feature subset FS of size $sizeFS$. Second, a graph G is created as just described using all (labeled and unlabeled) data points with features FS. G is used to predict the unlabeled data U through the "Label Propagation" algorithm. Next, we select the top $s\%$ unlabeled data with high confidence. The selected data points are added to the original training data with their labels, called PU. Thus, we have a new training dataset $newL = L + PU$. A newly improved feature subset $newFS$ is then determined from $newL$. Subsequently, a new graph is created from L, U and $newFS$. We repeat this process for $Iterations$ times and record average prediction confidence on the selected unlabeled data PU at each iteration. Finally, we select the feature subset with the highest average prediction confidence as the chosen feature subset $ResultFS$. The proposed approach is summarized in Algorithm 1. Time complexity of IteraGraph_FS can be obtained as follows. m is the number of features and n is the number of instances. Feature selection is bounded by $O(mn^2)$. The construction of graph G is bounded $O(mn^2)$, while "Label Propagation" is bounded by $O(n^3)$. The costs for calculating PU, $newFS(i)$, and $AvgConf(i)$ are bounded by $O(m \log m)$, $O(mn^2)$, and $O(n)$, respectively. These operations need to be performed $Iterations$ times. Therefore, the overall time complexity of IteraGraph_FS is bounded by $O((mn^2 + n^3) * Iterations)$.

3 Formal Analysis

First, we show that under a broad set of conditions (much broader than **i.i.d.** dimensions), as dimensionality increases the Euclidean distance from a data point x to its nearest neighbor approaches the Euclidean distance to x's farthest neighbor. We then provide some definitions that enable us to compare different distance measures. Based on these definitions, we analyze how a distance measure affects feature selection. Specifically, we show how to estimate classification confidence by the label propagation algorithm. We then establish the relationship between a distance measure and the effectiveness of confidence estimation. Finally, we justify our confidence-based sampling strategy, and thus make it clear how a distance measure affects the feature selection process, which depends on the sampled unlabeled data.

Definition 3.1 *Let $Q_m \in \mathbb{R}^m$. Let $\{N_{m,i}\}_{i=1}^n$ be n independent instances that are Q_m's neighbors. Let d_m be a function that calculates the Euclidean distance between Q_m and $N_{m,i}$. The* min *function and* max *function are*

$$min(D) = \min\{d_m(N_{m,i}, Q_m) \mid 1 \leq i \leq n\} \tag{4}$$

$$max(D) = \max\{d_m(N_{m,i}, Q_m) \mid 1 \leq i \leq n\} \tag{5}$$

Lemma 3.1 *If B_1,B_2,$\cdots$ is a sequence of random variables with finite variance such that $\lim_{m\to\infty} E[B_m] = b$ and $\lim_{m\to\infty} var(B_m) = 0$, then $\lim_{m\to\infty} P[B_m = b] = 1$.*

We have the following theorem which shows that the difference in distance between a point and all its neighbors becomes negligible as dimension increases.

Theorem 3.1 *Under the conditions in Definition 3.1, if*

$$\lim_{m\to\infty} var\left(\frac{(d_m(N_{m,1}, Q_m))^p}{E(d_m(N_{m,1}, Q_m)^p)}\right) = 0 \tag{6}$$

where $0 < p < \infty$ is a constant. Then $\forall \varepsilon > 0$

$$\lim_{m\to\infty} P[\max(D) \leq (1+\varepsilon)\min(D)] = 1 \tag{7}$$

Proof is available at [1]. Theorem 3.1 implies that the distance between any two examples is approximately the same under the stated conditions. Under these circumstances, it is difficult to tell to which class a data point belongs. Furthermore, similarity is a monotonic function of the distance, so is the "similarity measure". Because our algorithm relies on similarity between two examples, a "bad" similarity measure will be less useful to choose the best features. To explain how similarity affects the proposed algorithm, we first give some notations and definitions. The normalized similarity matrix is defined as $D^{-1}W$ with entries w_{ij}/d_i where $D = diag\{d_1, \ldots, d_{\ell+u}\}$. That is, the normalized similarity between two examples x_1 and x_2 is the similarity between x_1 and x_2 divided by the sum of the similarity between x_1 and all its neighbors.

Definition 3.2 Distinguishable normalized similarity measure *Let $S_1(\cdot,\cdot)$ and $S_2(\cdot,\cdot)$ be normalized similarity measures. $S_1(\cdot,\cdot)$ is more distinguishable than $S_2(\cdot,\cdot)$ $\iff$ Given an instance x_0 and $X = \{x_1, \cdots, x_n\}$. For any $x_k \in X$, $S_1(x_0, x_k) > S_2(x_0, x_k)$ if x_k has the same label as x_0 and $S_1(x_0, x_k) < S_2(x_0, x_k)$ otherwise.*

The unlabeled data sampling strategy depends on confidence. Thus a similarity measure that produces more consistent confidence estimates is preferred. Next, we explain how the distance measure affects confidence estimating through label propagation. Using the harmonic function f defined in Section2, we assign label 1 to an instance x_i if $f(i) > 0.5$ and label 0 otherwise (assuming a two class problem without loss of generality) [14]. Thus $x = 0.5$

represent the decision boundary under 0-1 loss. In terms of random walk, starting from an unlabeled node i, $f(i)$ can be viewed as the probability of hitting a labeled node with label 1, while $1 - f(i)$ can be viewed as the probability of hitting a node with label 0. Next we define the confidence.

Definition 3.3 *The classification confidence of instance x_i being labeled y_i using the label propagation algorithm is*

$$conf(y_i|x_i) = \frac{|f(i) - 0.5|}{0.5} \tag{8}$$

The above definition is consistent with the fact that the closer $f(i)$ is to the decision boundary 0.5, the less confident the classification result.

Then we can present the following theorem that provides a basis for using a more distinguishable distance measure.

Theorem 3.2 *With more distinguishable normalized similarity measure, a better classification confidence matrix f can be achieved in the sense that: 1. For a correctly classified instances x, the classification confidence will become larger. 2. For an instance x being classified incorrectly, the classification confidence will become smaller or the incorrect label will be corrected.*

Proof Because a node in the graph represents an instance, "node" j and "instance" x_j will be used exchangeably in this proof. Given a node x_j in the graph, let $X = \{x_1, \ldots, x_n\} = X_1 \cap X_2$ be the set of j's neighbors, where all the nodes in X_1 have the same label as x_j, and all the nodes in X_2 have a different label. Thus, $S_1(x_j, x_i) > S_2(x_j, x_i)$ and $S_1(x_j, x_k) < S_2(x_j, x_k)$, for every $x_i \in X_1$ and $x_k \in X_2$. Note that $w_{ij}/d_i = S(x_i, x_j)$, $\forall i, j$. For convenience, let $p^1_{ij} = S_1(x_i, x_j) = w^1_{ij}/d^1_i$ and $p^2_{ij} = S_2(x_i, x_j) = w^2_{ij}/d^2_i$. Without loss of generality, assume that node x_j belongs to class y_1. Note that for x_j, $\sum_{x_i \in X} p_{ij} = 1$. By the definition of X_1 and X_2, and according to how the harmonic function classifies examples, we have $f(i) > f(k)$, $x_i \in X_1$ and $x_k \in X_2$. Because $S_1(\cdot,\cdot)$ is more distinguishable than $S_2(\cdot,\cdot)$, $p^1_{ji} > p^2_{ji}$ and $p^1_{jk} < p^2_{jk}$. So long as there exists some $x_i \in X$ such that $f(i) \neq 0.5$, we have

$$\begin{aligned} f^1(j) - \frac{1}{2} &= \sum_{x_i \in X_1} p^1_{ji}[f(i) - \frac{1}{2}] + \sum_{x_k \in X_2} p^1_{jk}[f(k) - \frac{1}{2}] \\ &> \sum_{x_i \in X_1} p^2_{ji}[f(i) - \frac{1}{2}] + \sum_{x_k \in X_2} p^2_{jk}[f(k) - \frac{1}{2}] = f^2(j) - \frac{1}{2} \end{aligned}$$

The ">" in (9) can be substituted by "=" if and only if for all $x_i \in X$ such that $f(i) = 0.5$. If $f^2(j) > \frac{1}{2}$, then $f^1(j)$ should be greater than $\frac{1}{2}$ and farther from $\frac{1}{2}$ than $f^2(j)$. Thus we can obtain a more confident result under $S_1(\cdot,\cdot)$. On the other hand, if $f^2(j) < \frac{1}{2}$, there are two situations. If $f^1(j) > \frac{1}{2}$, then the classification is corrected. Otherwise, we have $f^2(j) < f^1(j) < \frac{1}{2}$, and the result become less confident under $S_2(\cdot,\cdot)$. □

Let x be an instance. We define the *near hit* or *nh* of x as its nearest neighbor that comes from the same class as x. Similarly, we define the *near miss* or *nm* as the nearest neighbor of x that comes from the opposite class. Then the hypothesis margin of x with respect to labeled data L is defined as $\theta(x) = \|x - nm(x)\| - \|x - nh(x)\|$ [4]. The hypothesis margin is easy to compute and lower bounds the sample margin. Notice that $\|\cdot\|$ is 1-Lipschitz. That is, $|\ \|x - nm(x)\| - \|x - nh(x)\|\ | \leq \|nm(x) - nh(x)\|$. Thus, we define the normalized hypothesis margin as

$$\sigma(x) = \frac{|\ \|x - nm(x)\| - \|x - nh(x)\|\ |}{\|nm(x) - nh(x)\|}. \tag{9}$$

This removes any differences due to scaling.

We now establish a relationship between $conf(x)$ and $\sigma(x)$ through the following lemma. It states that a higher confidence $conf(x)$ implies a larger margin $\sigma(x)$. We note that $f(x)$ (Eq. 2) is the harmonic average of its neighbors. However, our analysis employs the one nearest neighbor rule, i.e. the near miss or near hit in estimating $f(x)$. This simplification can be justified by noting that $P_{ul}f_l$ is the most dominant term in Eq. 3. This is also practiced in the analysis of Relief and Simba [4]. Nonetheless, our analysis explains the theoretical basis of our proposal.

Lemma 3.2 *Let x be an instance. The following hold: (1) An increase in $conf(x)$ results in an increase in $\sigma(x)$; and (2) A decrease in $conf(x)$ results in a decrease in $\sigma(x)$.*

Proof Assume x has a label +1. Let $w_{xnm(x)}$ and $w_{xnh(x)}$ be the weights along the edges from x to $nm(x)$ and from x to $nh(x)$ in the graph, respectively. There are two scenarios. First, $conf(x)$ increases. Since $conf(x)$ (Eq. 8) is monotonic in $f(x)$ (assuming $f(x) \geq 0.5$), an increase in $conf(x)$ results in an increase in $f(x)$. Now $f(x) = \frac{w_{xnh(x)}}{w_{xnm(x)} + w_{xnh(x)}} \times 1 + \frac{w_{xnm(x)}}{w_{xnm(x)} + w_{xnh(x)}} \times 0$, which follows from the first term in Eq. (3). Then, $f'(x) - f(x) > 0$ implies that $\frac{w'_{xnh(x)}}{w'_{xnm(x)} + w'_{xnh(x)}} > \frac{w_{xnh(x)}}{w_{xnm(x)} + w_{xnh(x)}}$. Consider two cases. Case 1: $w'_{xnm(x)} < w_{xnm(x)}$. Since Eq. (1) is monotonic in $\|x - nm(x)\|$, thus $\|x - nm'(x)\| > \|x - nm(x)\|$. This implies an increase in the margin $\sigma(x)$ (Eq. (9)). The proof of Case 2 ($w'_{xnh(x)} > w_{xnh(x)}$) is similar. The same argument can be applied to the second scenario. □

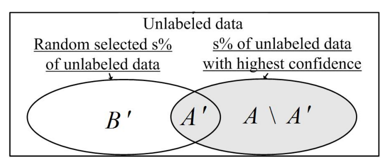


Figure 3. Illustration for Theorem 3.3

Next we show that our confidence-based unlabeled data sampling strategy provides an increased average margin over random selection.

Theorem 3.3 *Let $\bar{\sigma}(X) = \frac{\sum_{x_i \in X} \sigma(x_i)}{|X|}$ be the average margin of set X. Also, let A_{conf} and A_{rand} be the sets of the unlabeled data selected by our confidence-based strategy and the random strategy, respectively. Then the following holds:*

$$\bar{\sigma}(A_{conf}) \geq \bar{\sigma}(A_{rand}).$$

Proof A_{conf} consists of the top $s\%$ unlabeled data with the highest confidence. For random selection, the selected $s\%$ unlabeled data $A_{rand} = A'_{conf} \cup A'_{rand}$, where $A'_{conf} \subseteq A_{conf}$, and $A'_{rand} \cap A_{conf} = \phi$. Note that $|A_{conf}| = |A_{rand}|$. Clearly, $conf(x) \geq$

Table 1. Dataset summary

Dataset	$\lvert L\rvert$	$\lvert U\rvert$	Testing	Features
Biomedical and Gene Expression Data (HDLSS)				
Arcene	8	20	72	10000
Leukemia	8	20	44	7129
CNS	8	20	32	7129
Text Documents				
OrgsPeople	12	500	2445	9729
OrgsPlaces	12	500	2059	9729
PeoplePlaces	12	500	2154	9729

Table 2. Means and SD of Accuracy (%)

Dataset	Method	NaiveBayes	k-NN	SMO
Biomedical and Gene Expression Data (HDLSS)				
Arcene	SFFS	53.9±2.2	51.0±3.3	58.8±6.2
	sSelect	61.5±4.1	50.2±3.0	60.1±3.5
	IG_FS	**66.4**±9.0	**66.1**±5.9	**66.6**±7.4
Leukemia	SFFS	42.4±2.9	25.5±0.0	58.9±4.8
	sSelect	75.0±4.6	25.5±0.0	60.6±3.1
	IG_FS	**86.0**±11.6	**81.6**±12.5	**81.6**±9.7
Central Nervous System	SFFS	63.9±9.7	74.2±0.0	56.8±3.7
	sSelect	63.7±4.7	74.2±0.0	57.8±4.9
	IG_FS	**73.9**±2.9	**75.1**±1.3	**70.8**±5.6
Text Documents				
OrgsPeople	SFFS	61.7±0.0	61.7±0.0	57.9±0.0
	sSelect	52.4±0.7	52.4±0.0	52.4±0.0
	IG_FS	**64.0**±1.4	**62.7**±2.4	**58.6**±0.5
OrgsPlaces	SFFS	51.1±0.1	41.0±0.2	45.2±0.0
	sSelect	42.9±6.4	42.9±6.4	45.3±8.3
	IG_FS	**60.0**±0.7	**59.4**±0.0	**59.5**±0.1
PeoplePlaces	SFFS	38.8±0.0	35.8±0.0	36.5±0.0
	sSelect	51.1±12.2	57.1±9.6	51.1±12.3
	IG_FS	**61.3**±0.1	**61.7**±0.0	**61.7**±0.1

$conf(y)$, $\forall x \in A_{conf}$ and $\forall y \in A'_{rand}$. From Lemma 3.2, we have $\sigma(x) \geq \sigma(y)$, $\forall x \in A_{conf}$ and $\forall y \in A'_{rand}$. Therefore,

$$\begin{aligned}\bar{\sigma}(A_{conf}) &= \frac{1}{|A_{conf}|}\Big[\sum_{x_i\in A_{conf}\setminus A'_{conf}} \sigma(x_i) + \sum_{x_i\in A'_{conf}} \sigma(x_i)\Big] \\ &\geq \frac{1}{|A_{conf}|}\Big[\sum_{x_i\in A'_{rand}} \sigma(x_i) + \sum_{x_i\in A'_{conf}} \sigma(x_i)\Big] \\ &= \frac{1}{|A_{rand}|}\sum_{x_i\in A_{rand}} \sigma(x_i) = \bar{\sigma}(A_{rand}) \qquad (10)\end{aligned}$$

We obtain the stated result. □

Based on the above analysis, our framework should benefit from a better distance measure obtained from feature selection at earlier iterations, and from our sampling strategy that provides an improved margin over random selection. Our experimental results corroborate the above analysis.

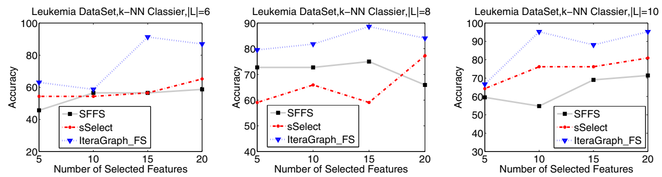
4 Experiment

The performance of IteraGraph_FS is compared with both traditional supervised method SFFS [8], and the state-of-art semi-supervised selection algorithm sSelect [12]. Three sets of studies are also conducted to further examine the sensitivity of the proposed method with respect to varied sizes of labeled or unlabeled training sets, and different percentages of selected unlabeled data.

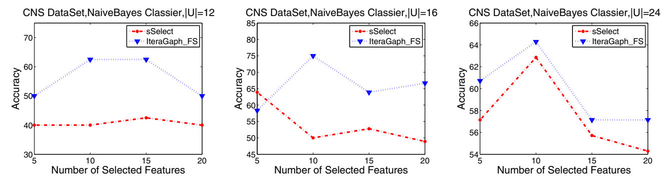
DataSet Description and Algorithm Setup As summarized in Table 1, two types of binary datasets are used in the study, including biomedical and gene expression data [6], and text documents[7]. For each dataset, the labeled (L) and unlabeled instances (U) are randomly selected, and the rest are used for the testing purpose. Importantly, these three biomedical datasets are typical examples of high dimensional, low sample size (HDLSS) problems in which dimension reduction through feature selection is inevitable. To execute IteraGraph_FS, we set the parameters *Iterations* and s to be 10 and 90% respectively. For sSelect and IteraGraph_FS, both the labeled and unlabeled data points are utilized to perform semi-supervised feature selection; While SFFS approach can only be trained on the labeled instances. After selecting the feature subset FS of a specific size $sizeFS$, we construct the classifier only with the labeled instances and FS, and then the unseen testing instances are employed to evaluate the classification accuracy. Several popular algorthims such as Naive Bayes, kNN and SMO are used as the subsequent classifiers, and the algorithm implementations are based on Weka.

Feature Quality Study To fully investigate the overall discriminative ability of selected features acquired by SFFS, sSelect and IteraGraph_FS as the size of chosen features changes, we varied the value of $sizeFS$ from 5 to 20 for each dataset. The means and standard deviations of the accuracies over 16 different feature subset sizes are listed in Table 2. Remarkably, as highlighted in bold, the means of the arracuracies of IteraGraph_FS are higher than those of the two others in all comparisons. On every dataset involved in the study, the means of IteraGraph_FS are the highest ones for the three learning algorithms, along with relatively small standard deviations. The better performance of IteraGraph_FS over sSelect could be ascribed to the graph-based distance measure implemented in IteraGraph_FS, which allows more distinguishable distance estimation than the traditional Euclidean distance due to the curve of dimensionality [1]. On the other hand, the worst accuracy of SFFS is due to a rather small number of labeled training instances, which is the inherent characteristic of HDLSS datasets. The similar performance explanation can be applied to other datasets.

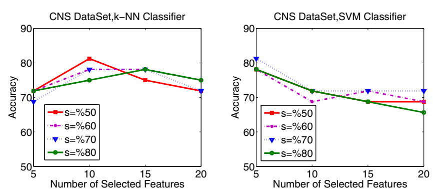
Algorithmic Sensitivity Studies We also studied the sensitivity of IteraGraph_FS from the following three aspects: different sizes of $|L|$, different sizes of $|U|$, and varied percentages of s. Firstly, for different size of $|L|$, as shown in the Fig.4(a), IteraGraph_FS consistently outperforms or performs equally as sSelect and SPPS. In general, for the three methods, the more labeled data are used, the better accuracy can be achieved. However, the performance of IteraGraph_FS does not significantly improve as the size of the labeled training set increases from 6 to 10. This implies that IteraGraph_FS could be able to select the features of high discriminability even with a small number of labeled training points. Secondly, for different size of $|U|$, it can be observed from Fig.4(b) that IteraGraph_FS consistently outperforms sSelect. A closer examination reveals that, in general, the accuracy of IteraGraph_FS does not strictly increase as more unlabeled data are involved. This suggests that, although unlabeled data can provide additional information, in the meanwhile, they also inevitably introduce more



(a) Accuracy (%) vs. Different $|L|$



(b) Accuracy (%) vs. Different $|U|$



(c) Accuracy (%) vs. Different s

Figure 4. Studies of Algorithmic Sensitivity

uncertainty. Finally, as the values of s vary from 50% to 80%, Fig.4(c) shows that, for both learners, the span of the accuracy fluctuation is less than 5%, demonstrating that IteraGraph_FS is not influenced excessively by different values of s.

5 Related Works

Many supervised and un-supervised feature selection methods have been proposed, such as (but not limited to) [3, 10, 13]. However, semi-supervised feature selection has most recently received a lot of interests [12, 5, 11]. Recently, [12] presents a semi-supervised feature selection algorithm based on spectral analysis that exploits both labeled and unlabeled data through a regularization framework. However, it uses all features to construct the graph, thus making the distance measure non distinctive and its performance can be unsatisfactory at times. Most recently, [2] proposes a graph classifier based on kernel smoothing method aiming to optimize certain loss functions. [9] discusses a new label propagation algorithm to address the label imbalances problem, which could be adopted by our approach to handle similar dataset.

6 Conclusion

We have introduced and explored a new concept of "hybrid feature selection" using both supervised and semi-supervised methods. Instead of using labeled and unlabeled data at the same time, the proposed approach works in an iterative procedure that starts with labeled information to select some critical features, and then uses the unlabeled data to improve this chosen feature set by either excluding some features or including new ones. We use a graph model to explore the use of unlabeled data to measure each feature's separability. The quality of chosen features by the proposed approach is analyzed by demonstrating both better distance measure, as well as higher prediction confidence on unlabeled examples. Empirically, we have shown that when the size of the labeled data is small ($\times 10$) and the number of features is large ($\times 10^3$), the proposed approach can select features with at least 10% higher in accuracy than both a known supervised method and a recently proposed semi-supervised feature selection algorithm.

References

[1] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? *Lecture Notes in Computer Science*, 1540:217–235, 1999.

[2] M. Culp and M. G. Graph-based semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):174–179, 2008.

[3] J. Dy and C. E. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889, 2005.

[4] R. Gilad-Bachrach, A. Navot, and N. Tishby. Margin based feature selection - theory and algorithms. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 43, New York, NY, USA, 2004. ACM.

[5] J. Handl and J. Knowles. Semi-supervised feature selection via multiobjective optimization. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2006)*, pages 6351–6358, 2006.

[6] L. Jinyan and L. Huiqing. Kent ridge bio-medical data set repository, 2005. http://leo.ugr.es/elvira/DBCRepository/index.html.

[7] D. D. Lewis. Reuters-21578 test collection, 2004. http://www.daviddlewis.com.

[8] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Journal of Pattern Recognition Letters*, 15:1119–1125, 1994.

[9] J. Wang, T. Jebara, and S. Chang. Graph transduction via alternating minimization. In A. McCallum and S. Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 1144–1151. Omnipress, 2008.

[10] A. Wolf, L. & Shashua. Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight-based approach. *Journal of Machine Learning Research*, 6:1855–1887, 2005.

[11] D. Zhang, Z.-H. Zhou, and S. Chen. Semi-supervised dimensionality reduction. In *SIAM International Conference on Data Mining*, 2007.

[12] H. Zhao, Z. & Liu. Semi-supervised feature selection via spectral analysis. In *SIAM International Conference on Data Mining*, 2007.

[13] H. Zhao, Z. & Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceeding of the 24th International Conference on Machine Learning*, pages 1151–1158, 2007.

[14] X. Zhu. *Semi-supervised learning with graphs*. PhD thesis, Dept. of Computer Science, University of Carnegie Mellon, 2005.