STABLE DYSPHONIA MEASURES SELECTION FOR PARKINSON SPEECH REHABILITATION VIA DIVERSITY REGULARIZED ENSEMBLE

Wei Ji[*]

College of Telecommunications
and Information Engineering
Nanjing University of Posts
and Telecommunications, Nanjing, P. R. China

Yun Li [†]

College of Compter Science
Nanjing University of Posts
and Telecommunications, Nanjing,
P. R. China

ABSTRACT

Vocal impairment is a common symptom for the vast majority of Parkinson's disease (PD) subjects. And it needs long term rehabilitation through personalized one-to-one periodic rehabilitation meetings with clinical speech experts. The significant challenge is that there are not enough experts to deliver the in-person treatments that is needed and for many people with PD, it is difficult to visit the experts for monitoring and treatments. Then there is the need for reliable clinical tools to assist the rehabilitation. This study aims to investigate the potential of using sustained vowel phonations towards objectively and automatically replicating the speech experts' assessments of PD subjects' voices as "acceptable" (a clinician would allow persisting during in-person rehabilitation treatment) or "unacceptable" (a clinician would not allow persisting during in-person rehabilitation treatment). The phonation is usually characterized by many dysphonia measures, which are extracted by clinical speech signal processing algorithms. For this aim, we need to select a stable dysphonia measures subset, and adopt it to automatically distinguish the PD subjects' voices (acceptable versus unacceptable). In this paper, a diversity regularized ensemble feature weighting algorithm DREFW is presented to choose the stable dysphonia measures subset. The experimental results on real speech rehabilitation data set have shown the proposed algorithm can obtain high stability and classification performance for speech assessment. The findings of this paper is a first step towards improving the effectiveness of an automated rehabilitative speech assessment tool.

Index Terms— Parkinson, Speech Rehabilitation, Dysphonia Measures, Feature Selection

1. INTRODUCTION

Parkinson's disease (PD) profoundly affect the lives of patients and their families. Especially, vocal impairment is reported in the vast majority of PD subjects, and approximately 29% of those consider it one of their greatest hindrances associated with the disease [1]. The extent of vocal impairment can be assessed using sustained vowel phonations [2, 3]. Then for the rehabilitative speech treatment, the clinical speech experts should assess the sustained vowel phonations as "acceptable (a clinician would allow persisting in speech treatment) or "unacceptable (a clinician would not allow persisting in speech treatment). Now, the significant challenge is the barrier of inadequate numbers of clinicians to deliver in-person therapy, enhancing the feasibility of delivering intensive treatment requirements, and relieving the logistical burden of travelling to and from the clinic for in-person treatment. Then it is very important to supply the convenient system to assist the speech rehabilitation. Advances in computer and web-based technology offer solutions to the problems of treatment accessibility, efficacious treatment delivery, and long-term maintenance in rehabilitation [4]. For the rehabilitation system, the current study is to investigate the potential of using an objective statistical machine learning framework to automatically evaluate sustained vowel phonations as "acceptable" or "unacceptable". Phonations are characterized by many dysphonia measures, which are extracted by clinical speech signal processing algorithms. In this framework, support vector machines (SVMs) [5] is preferred to classify the sustained vowel phonation as "acceptable" or "unacceptable" on the basis of dysphonia measures [1]. Moreover, in this framework, another key problem is to identify the stable and effective dysphonia measures subset for phonations evaluation, i.e., choose stable information-rich dysphonia measures, then map them to the response (acceptable versus unacceptable) through SVM [1]. The dysphonia measures selection can reduce the dimensionality of dysphonia measures space to alleviate the "curse of dimensionality" for classification and to improve the classification accuracy of phonations evalua-

[*]Natural Science Foundation of Jiangsu Province BK20140885, Postdoctoral Foundation of Jiangsu Province (1401045C) and Foundation of Key Laboratory of Cloud Computing and Complex System (15206).

[†]Corresponding Author. Natural Science Foundation of Jiangsu Province BK20131378 and NSFC 61302157, 61300164

tion. In addition, the chosen stable dysphonia measures retain domain expertise and they can be used to detect unacceptable voice characteristics during use of software away from expert clinical guidance, stop the patient from using voice in an unacceptable way, and subsequently improve voice characteristics through providing feedback.

The dysphonia measures selection can be considered as a feature selection problem in machine learning. According to the introduction above, the selection result should be effective and stable. Since local learning based feature selection has been shown high performance [6], and ensemble technique can be used to improve the robustness of feature selection [7, 8], then the ensemble local learning-based feature selection was adopted in [1] for dysphonia measures selection. In order to improve the performance of dysphonia measures selection for Parkinson speech rehabilitation, a diversity regularized ensemble feature weighting algorithm-DREFW is presented in this paper. The base feature selector in this ensemble is also based on local learning, moreover, the diversity between base selectors is considered in ensemble model. It is well-known that the generalization error of an ensemble is related to the average generalization error of the base learners and the diversity among the base learners. Generally, the lower the average generalization error (or, the higher the average accuracy) of the base learners and the higher the diversity among the base learners, the better the ensemble [9]. The experimental results show DREFW obtains higher identification performance for Parkinson speech rehabilitation in most cases without sacrificing the results stability.

2. DIVERSITY REGULARIZED ENSEMBLE FEATURE WEIGHTING-DREFW

2.1. Components of Ensemble Feature Selection

Ensemble feature selection firstly create a set of different base feature selectors, each provides its output (feature weighting vector or a feature subset), then aggregates the results of all base feature selectors to obtain the ensemble result [7]. In other words, ensemble feature selection consists of two components, i.e. the base feature selectors and the combination strategy of their output.

To produce the base feature selector, we adopt a subsampling based strategy. Consider a training set $\mathbf{X}$ contains *n* samples, $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, and each sample $\mathbf{x}_i$ is represented by an *d*-dimensional vector $\mathbf{x}_i \in \mathcal{R}^d$ and discrete class labels y_i. Then *m* subsamples of size $\beta n (0 < \beta < 1)$ are drawn randomly from $\mathbf{X}$, where the parameters *m* and β can be varied. Subsequently, feature selection is performed on each of the *m* subsamples to create the base feature selector. In our case, the feature weighting algorithm is utilized to produce base feature selector and its output is a feature weight vector for all features. Therefore, ensemble feature selection on *m* subsamples generates the feature weighting results set $\mathbf{E} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m\}$, where $\mathbf{w}_k (k = 1, 2, \cdots, m)$ represents the outcome of the *k*-th base feature selector trained on *k*-th subsample. The results set $\mathbf{E}$ can be obtained by minimizing the following loss function, which is based on the idea that maximize the fit of the feature weighting vector, while maximizing the diversity between vectors:

$$L(\mathbf{E}) = L_{emp}(\mathbf{E}) + \gamma . L_{div}(\mathbf{E}), \tag{1}$$

where $L_{emp}(\mathbf{E})$ is the empirical loss of $\mathbf{E}$; $L_{div}(\mathbf{E})$ is the diversity loss of $\mathbf{E}$ and it can be recognized as regularization term to embedding some prior knowledge. γ is the cost parameter balancing the importance of the two terms.

Since local-learning based feature weighting has shown to be efficient for dysphonia measures selection [1], we employ local learning-based logistic regression to implement the base feature selectors. Thus, the first term $L_{emp}(\mathbf{E})$ in Eqn.(1) is set to measure the empirical loss of logistic regression for feature weighting:

$$L_{emp}(\mathbf{E}) = \sum_{k=1}^{m} \sum_{x_i \in k^{th} subsample} \log(1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i)), \tag{2}$$

where $\mathbf{z}_i = |\mathbf{x}_i - N_{miss}(\mathbf{x}_i)| - |\mathbf{x}_i - N_{hit}(\mathbf{x}_i)|$, and $|.|$ is an element-wise absolute operator. $\mathbf{x}_i$ is a sample in *k*-th subsample. And two nearest neighbors of sample $\mathbf{x}_i$, one from the same class is called as nearest hit (N_{hit}), and the other from the different class is named as nearest miss (N_{miss}). $\mathbf{w}_k = (w_k^1, w_k^2, \cdots, w_k^d)$ is a vector of length *d* and $w_k^t (t = 1, 2, \cdots, d)$ represents the weight for feature *t* in *k*-th base feature selector output. $\mathbf{w}_k^T \mathbf{z}_i$ is the local margin for $\mathbf{x}_i$, which belongs to hypothesis margin [10] and an intuitive interpretation of this margin is to measure how much the features of $\mathbf{x}_i$ can be corrupted by noise (or how much $\mathbf{x}_i$ can "move" in the feature space) before being misclassified. The natural idea behind the Eqn. (2) is to obtain a weighted feature space parameterized by a feature weights vector $\mathbf{w}_k$. So that a margin-based error function in the induced feature space is minimized. For the purposes of this paper, we use the Manhattan distance to define the margin and nearest neighbors, while other standard distance definitions may also be used. Note that the defined margin only requires the information about the neighborhood of $\mathbf{x}_i$, while no assumption is made about the underlying data distribution. This means that we can transform an arbitrary nonlinear problem into a set of locally linear ones by local learning [6].

As shown in Eqn.(1), the regularization term $L_{div}(\mathbf{E})$ is used to characterize the diversity loss among the base feature selectors. Though there is no agreement on what form of diversity should be defined, the diversity measures usually can be defined in a pairwise form. Thus we consider a form of diversity based on pairwise difference, and then the form of diversity loss is defined as pairwise similarity. The more similar all outputs are, the higher the diversity loss measure will be. The overall diversity loss can be defined as the average over all pairwise similarity between the outputs of different base feature selectors:

$$L_{div}(\mathbf{E}) = \frac{1}{m(m-1)} \sum_{k=1}^{m-1} \sum_{k'=k+1}^{m} Sim(\mathbf{w}_k, \mathbf{w}_{k'}), \qquad (3)$$

where $Sim(\mathbf{w}_k, \mathbf{w}_{k'})$ represents a similarity measure between feature weighting vector $\mathbf{w}_k$ and $\mathbf{w}_{k'}$. Notice that the feature weighting vector is direct related to the classification error based on the margin as described above, and each feature weighting vector $\mathbf{w}_k$ is linear without the bias term, thus the direction of vector is the most important factor for the classification performance. In the paper, the cosine similarity measure is adopted with normalized feature weights to calculate the similarity between weighting vector $\mathbf{w}_k$ and $\mathbf{w}_{k'}$, then $Sim(\mathbf{w}_k, \mathbf{w}_{k'}) = \mathbf{w}_k^T \mathbf{w}_{k'}$. Note that the adding of a constant $||\mathbf{w}_k||_2^2 + ||\mathbf{w}_{k'}||_2^2$ (its value is 2) does not change the optimal solution [11]. In this case, the diversity loss can be replaced by $||\mathbf{w}_k + \mathbf{w}_{k'}||_2^2$, i.e.

$$L_{div}(\mathbf{E}) = \frac{1}{m(m-1)} \sum_{k=1}^{m-1} \sum_{k'=k+1}^{m} ||\mathbf{w}_k + \mathbf{w}_{k'}||_2^2, \qquad (4)$$

and a relaxed convex optimization problem is obtain for ensemble feature weighting loss in Eqn.(1). Furthermore, the diversity loss is also a *l*2-norm regularization term for logistic regression, which leads to the stable feature weighting vectors for its robustness to the rotational variation [12]. Then the proposed diversity loss term has positive effect on feature selection stability besides the classification performance.

In summary, ensemble feature selection aims to find the target model $\mathbf{E}^*$ through minimizing the loss function in Eqn.(1):

$$\mathbf{E}^* = argmin_{\mathbf{w}_k} L(\mathbf{E}), \qquad (5)$$

and the final ensemble feature weighting result is obtained by linear combination of the outputs of base feature selectors.

$$\mathbf{w}_e = \frac{1}{m} \sum_{k=1}^{m} \mathbf{w}_k, \qquad (6)$$

where $\mathbf{w}_k \in \mathbf{E}^*$.

The target model $\mathbf{E}^*$ can be found by employing gradient descent-based techniques. Accordingly, the gradients of $L(\mathbf{E})$ in Eqn. (1) w.r.t the model parameters $\Theta = \{\mathbf{w}_k | 1 \leq k \leq m\}$ are determined as follows:

$$\frac{\partial \mathbf{L}}{\partial \Theta} = [\frac{\partial \mathbf{L}}{\partial \mathbf{w}_1}, \cdots, \frac{\partial \mathbf{L}}{\partial \mathbf{w}_k}, \cdots, \frac{\partial \mathbf{L}}{\partial \mathbf{w}_m}], \qquad (7)$$

where

$$\begin{aligned} \frac{\partial \mathbf{L}}{\partial \mathbf{w}_k} &= \frac{1}{\beta n} \sum_{x_i \in k^{th} subsample} \frac{\partial \log(1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i))}{\partial \mathbf{w}_k} \\ &+ \frac{2\gamma}{m(m-1)} \sum_{k'=1, k' \neq k}^{m} \frac{\partial Sim(\mathbf{w}_k, \mathbf{w}_{k'})}{\partial \mathbf{w}_k}. \qquad (8) \end{aligned}$$

and

$$\frac{\partial \log(1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i))}{\partial \mathbf{w}_k} = -\frac{1}{m} \frac{\exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i)}{1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i)} \mathbf{z}_i,$$

$$\frac{\partial Sim(\mathbf{w}_k, \mathbf{w}_{k'})}{\partial \mathbf{w}_k} = 2(\mathbf{w}_{k'} + \mathbf{w}_k). \qquad (9)$$

Now, we are at the position to summarize the diversity regularized ensemble feature weighting algorithm DREFW in **Algorithm 1**.

Algorithm 1 The Diversity Regularized Ensemble Feature Weighting (DREFW) algorithm

Step 1. Input training data set $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$ and regularization parameter γ in Eqn. (1).
Step 2. Initialize $\mathbf{w}_k \in \mathbb{R}^d$ where $k = 1, \cdots, m$.
Step 3. For $k = 1, 2, \cdots, m$-th subsampling
  For every x_i in *k*-th subsample
    Minimizing Eqn. (1) through Eqns. (8) and (9) to obtain $\mathbf{w}_k \in \mathbf{E}^*$ ($k = 1, \cdots, m$)
Step 4. Output the ensemble feature weighting result using Eqn. (6).

To initialize the ensemble, each feature selector is learned from a bootstrapped sample of $\mathbf{X}$. Specifically, the corresponding feature weighting $\mathbf{w}_k$ is obtained by minimizing the objective function $\mathbf{w}_k = \min_{\mathbf{w}_k} \sum_{x_i \in k^{th} subsample} \log(1 + \exp(-\mathbf{w}_k^T \mathbf{z}_i))$. Note that the ensemble can also be initialized in other ways, such as instantiating each $\mathbf{w}_k$ with random values, etc.

2.2. Stability Analysis

Same to the steps introduced in [7, 13, 14], the stability of ensemble feature weighting algorithm is calculated as follows: Consider the data set ***S*** with *Q* instances and *d* features. Then *c* sample subsets of size $\mu Q (0 < \mu < 1)$ are drawn randomly from ***S***, where the parameters *c* and μ also can be varied. The sample subset is used as $\mathbf{X}$ described above. Subsequently, ensemble feature weighting is performed on each of the *c* sample subsets, and the similarity of outputs of ensemble feature weighting on the *c* sample subsets are calculated. The more similar all outputs are, the higher the stability will be. The overall stability can be defined as the average similarity over all pairwise similarity between the different ensemble feature weighting results. However, feature weighting is almost never directly used to compute the stability of feature selection, and instead converted to a ranking result based on the weights. For feature ranking, the Spearman rank correlation coefficient [7, 15] can be used to calculate the similarity.

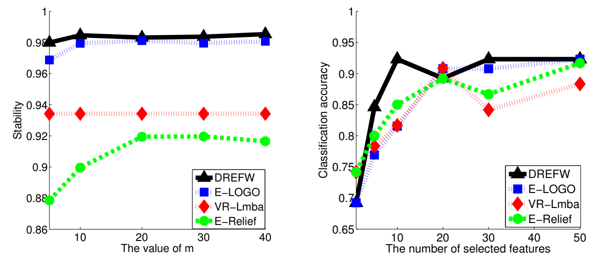
3. EXPERIMENTS

3.1. Data

The experimental Parkinson speech rehabilitation data set is from [1], which is derived from 14 PD subjects (eight males and six females) with an age range of 51-69 years, and produced by sustained vowel /a/ phonations. The phonations were assessed by experts perceptually whether phonations could be "acceptable" or "unacceptable". The dysphonia measures used to characterize the phonation are defined and summarized in [16]. Refer to [1, 16], we simply summarize the dysphonia measures and cluster them into groups. The first group of dysphonia measures builds on the physiological observation that the vocal fold vibration pattern is nearly periodic in healthy voices, whereas pathological voices tend to depart from periodicity or are completely aperiodic. Two of the most widely used dysphonia measures fall under this category are known as jitter and shimmer. The recurrence period density entropy (RPDE), the pitch period entropy (PPE) and the glottal quotient (GQ) are also fallen into the first group. The second general group of dysphonia measures is the signal-to-noise ratio (SNR) type algorithms. The physiological motivation for this group is that incomplete vocal fold closure leads to the creation of aerodynamic vortices which result in increased acoustic noise. Harmonic-to-noise ratio (HNR), detrended fluctuation analysis (DFA), glottal to noise excitation (GNE), vocal fold excitation ratio (VFER), and empirical mode decomposition excitation ratio (EMD-ER) are archetypal examples of this group. Lastly, Mel frequency cepstral coefficients (MFCCs) target the placement of the articulators (collectively referring to the mouth, teeth, tongue, and lips), which is known to be affected in PD. Overall, 309 dysphonia measures are calculated to describe each phonation, resulting in a design matrix of size 156×309.

3.2. Experimental Results for Stability

In this part of experiments, we will validate the stability of proposed ensemble feature weighting algorithm DREFW and compare with other state-of-the-art stable algorithms, such as ensemble-LOGO (E-LOGO) used in [1], ensemble-Relief (E-Relief) [7] and VR-Lmba [13, 14]. To estimate the stability of ensemble feature weighting algorithm, the strategy explained above was used with $c = 5$ sample subsets of size 0.9Q (i.e. $\mu = 0.9$ and each sample subset contains 90% of the data). In our case, the size of sample subset is $0.9 \times 156 = 140$. This percentage was chosen because we want to assess stability with respect to relatively small changes in the data set. Then, the proposed ensemble algorithm with $\beta = 0.9$ was run on each sample subset, and stability is calculated as described in section 2.2. We show the stability of these algorithms w.r.t different numbers of base feature selectors, i.e. the value of m, in Fig.1(a). We see that the stability of all algorithms saturates at around m=20. Since VR-Lmba is not an ensemble method, its stability remains constant.



(a) Stability (b) Accuracy

Fig. 1. Experimental results of stability and accuracy

3.3. Experimental Results for Classification

Now we will turn to validate the classification performance of our proposed DREFW for Parkinson speech rehabilitation. In this part of experiments, the number of base selectors for ensemble feature weighting is constant and set as 20 for all ensemble algorithms, i.e., m=20. 10-cross validation is used and the linear SVM is adopted as classifier with C=1 [17]. The accuracy rate of different numbers of selected dysphonia measures for classifying the patients' phonation are shown in Fig. 1(b). The top ten dysphonia measures selected by DREFW is $\{GNE_{NSR,TKEO}$, $VFER_{SNR,SEO}$, $VFER_{NSR,TKEO1}$, $VFER_{NSR,TKEO}$,$VFER_{NSR,SEO}$, $IMF_{NSR,TKEO}$, *Log energy*, $0^{th}MFCC$, $2^{nd}MFCC$, $5^{th}MFCC\}$

From the experimental results above, we can observe that our proposed ensemble algorithm-DREFW, can obtain higher classification accuracy than other ones in most cases. The stability value of our algorithm is approaching to 1 and is superior or at least equivalent to other methods. Then the diversity regularization term in our proposed ensemble feature weighting algorithm is effective to improve the classification performance without sacrificing the stability.

4. CONCLUSION

To stably and effectively choose the dysphonia measures for automatically Parkinson's speech rehabilitation, a diversity regularized ensemble feature weighting algorithm-DREFW is presented. Local learning-based base feature selector is adopted and diversity between base selectors is considered in the evaluation criterion. The DREFW is applied into the Parkinson's speech rehabilitation to find the stable information-rich dysphonia measures, and combined with SVM to classify the sustained vowel phonation as "acceptable" or "unacceptable". The experimental results have shown its higher accuracy and at least similar stability to other ones.

5. REFERENCES

[1] A. Tsanas, M. A. Little, C. FOX, and L. O. Ramig, “Objective automatic assessment of rehabilitatative speech treatment in parkionson’s disease,” *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 22, pp. 181–190, 2014.

[2] V. Parsa and D. G. Jamieson, “Acoustic discrimination of pathological voice: sustained vowels versus continuous speech,” *Journal of Speech, Language, and Hearing Research*, vol. 44, pp. 327–339, 2001.

[3] M. A. Little, D. A. E. Costello, and M. L. Harries, “Objective dysphonia quantification in vocal fold paralysis: comparing nonlinear with classical measures,” *Journal of Voice*, vol. 25, pp. 21–31, 2009.

[4] E. Taub, P. S. Lum, P. Hardin, B. W. Mark, and G. Uswatte, “Autocite: Automated delivery of ci therapy with reduced effort by therapists,” *Stroke*, vol. 36, pp. 1301–1304, 2005.

[5] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.

[6] Y. J. Sun, S. Todorovic, and S. Goodison, “Local learning based feature selection for high dimensional data analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1–18, 2010.

[7] Y. Saeys, T. Abeel, and Y. V. de Peer, “Robust feature selection using ensemble feature selection techniques,” in *Proceedings of the 25th European Conference on Machine Learning and Knowledge Discovery in Databases*, Banff, Canada, 2008, pp. 313–325.

[8] S. Loscalzo, L. Yu, and C. Ding, “Consensus group stable feature selection,” in *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009, pp. 567–575.

[9] A. Krogh and J. Vedelsby, “Neural network ensembles, cross validation, and active learning,” in *Advances in Neural Information Processing Systems*, 1995, pp. 231–238.

[10] K. Crammer, R. G. Bachrach, A. Navot, and N. Tishby, “Margin analysis of the lvq algorithm,” in *Advances in Neural Information Processing Systems*, 2002, pp. 462–469.

[11] Y. Yu, Y. F. Li, and Z. H. Zhou, “Diversity regularized machine,” in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 2011, pp. 1603–1608.

[12] A. Y. Ng, “Feature selection, l1 vs. l2 regularization, and rotational invariance,” in *Proceedings of International Conference on Machine Learning*, Banff, Canada, 2004.

[13] Y. Li, S. Y. Gao, and S. C. Chen, “Ensemble feature weighting based on local learning and diversity,” in *AAAI Conference on Artificial Intelligence*, 2012, pp. 1019–1025.

[14] Y. Li, J. Si, G. Zhou, S. Huang, and S. Chen, “Frel:a stable feature selection algorithm,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, pp. 1388–1402, 2015.

[15] A. Kalousis, J. Prados, and M. Hilario, “Stability of feature selection algorithms: a study on high dimensional spaces,” *Knowledge and Information Systems*, vol. 12, pp. 95–116, 2007.

[16] A. Tsanas, “Accurate telemonitoring of parkinsons disease symptom severity using nonlinear speech signal processing and statistical machine larning,” Ph.D. dissertation, University of Oxford, 2012.

[17] C. C. Chang and C. J. Lin, “Libsvm: a library for support vector machines,” in *http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.ps.gz*, 2002.