Power SVM: Generalization with Exemplar Classification Uncertainty

Weiyu Zhang
University of Pennsylvania
zhweiyu@seas.upenn.edu

Stella X. Yu
UC Berkeley / ICSI
stellayu@cs.berkeley.edu

Shang-Hua Teng
University of Southern California
shanghua@usc.edu

Abstract

The human vision tends to recognize more variants of a distinctive exemplar. This observation suggests that discriminative power of training exemplars could be utilized for shaping a desirable global classifier that generalizes maximally from a few exemplars. We propose to derive classification uncertainty for each exemplar, using a local classification task to separate the exemplar from those in other categories. We then design a global classifier by incorporating these uncertainties into constraints on the classifier margins. We show through the dual form that the classification criterion can be interpreted as finding closest points between convex hulls in the feature space augmented by classification uncertainty. We call this scheme Power SVM (as in Power Diagram), since each exemplar is no longer a singular point in the feature space, but a super-point with its own governing power in the classifier space. We test Power SVM on digit recognition, indoor-outdoor categorization, and large-scale scene classification tasks. It shows consistent improvement over SVM and uncertainty weighted SVM, especially when the number of training exemplars is small.

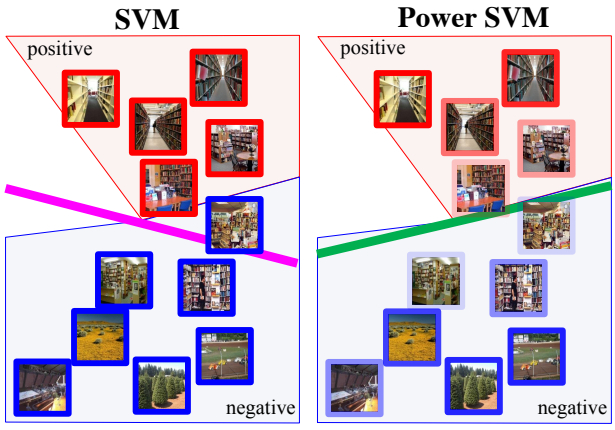

Figure 1. Power SVM utilizes different discriminative powers of exemplars to train a classifier that generalizes better from a few exemplars, while conventional SVM considers exemplars as homogeneous points in the feature space where only their locations matter. For example, to separate (*Library Indoor*) scene category from others, long aisle views of bookshelves (redder outlines) would be less confusing than front views of bookshelves and should lie farther away from the decision boundary of any desirable classifier.

1. Introduction

We study the problem of discriminating *many* classes with *limited* training data. This is a practical problem particularly for scene categorization, where the distribution of objects in semantic categories follows a power law [8] and infrequent objects have very few labeled data.

Our basic observation is that exemplars have different discrimination capacities, and a distinctive exemplar allows our human vision to recognize more of its variants. In computer vision, we can evaluate how easily an exemplar might be confused with others, and use such exemplar-centric local discrimination information to constrain a global classifier: A good classifier should place less confusing exemplars farther away from the global decision boundary.

In the max margin classification framework, we develop a model called *Power SVM* (Fig.1), where each exemplar is no longer just a singular point in the data space as in conventional SVM, but a super-point with certain prescribed governing power in the classifier space. This concept of non-uniform uncertainty among exemplars is similar to that of Power Diagram, a generalization of Voronoi Diagram where singular points become balls with distinctive radii [7].

By acknowledging the distinction in discrimination capacity among exemplars, we in fact enhance the utility of each exemplar in learning a global classifier that generalizes better from the few exemplars. The fundamental problem of learning is not just to associate a specific data example with its conceptual class, but to transfer that association beyond apparent resemblance between unseen test data and training exemplars. Power SVM is a simple yet effective and universal approach towards this goal.

There have been several lines of efforts on multiclass learning from limited labelled data. Semi-supervised learning [31, 11] propagates label information to vast unlabeled data. It requires less labeling, but the effectiveness relies on the local density of massive unlabeled data. On the other hand, active learning [15, 27] selects informative exemplars

and not all the available data are used. Hierarchical approaches [23, 1, 12] introduce a tree structure to classify multiple categories efficiently. A binary classification problem is solved at each decision node, with target categories grouped into positive and negative sets based on their categorical confusion. Exemplars in the same category are still treated in a homogeneous fashion.

Several approaches also introduce priors on model parameters, context, or human performance data, in order to increase the generalization capability of a few exemplars [24, 10, 25]. These priors are often feature-classifier specific and require knowledge outside the training data.

Our work distinguishes itself in three aspects: 1) We obtain classification uncertainty information based entirely on labelled data; 2) We differentiate exemplars in the same category based on their levels of confusion towards other categories; 3) We directly evaluate exemplars' discriminative power in the same feature space for the final classification.

Our work can be compared with several SVM methods. While structured SVM also introduces bias among exemplars [5], their bias is acquired from rich labeling information (e.g. object mask) that we do not require. While SVM with nearest neighbours approaches learn a discriminative distance function for each point in the feature space [9, 29], we evaluate the discrimination capacity of each exemplar and formulate that into a constraint on the classifier.

Compared to curriculum learning methods [2, 16] which also utilize the quality of individual exemplars, our method is not iterative and does not require any initialization: It evaluates all the exemplar uncertainty at once in order to learn a global classifier.

Our exemplar-centric view of classification is related to many non-parametric approaches in scene recognition [26, 13], scene parsing [20, 19] and object recognition [30, 22]. All these methods often work well only with a vast number of exemplars. While we train a local classifier for each exemplar, we do not compare a test image to all the exemplars according to their local classifiers [22]. We use local classifiers' performance on the training data to shape a global classifier, achieving a faster and better test performance that would otherwise require a lot more training data.

2. Power SVM

Consider training a binary linear classifier that separates m positive exemplars from n negative exemplars. We order the exemplars so that the first m belong to the positive class and the last n belong to the negative class.

Each exemplar is specified by (x_i, y_i, u_i), where x_i is a d-dimensional feature vector, $y_i \in \{-1, +1\}$ is the binary class label, and u_i is a non-negative uncertainty value indicating how easily the exemplar with feature x_i might be confused with those from the opposite class. The easier the confusion, the larger the uncertainty. Intuitively, if the exemplar takes a feature value between the positive and negative classes, there is maximal classification uncertainty, whereas if it takes a typical feature value of the class, then $u_i = 0$, there is no classification uncertainty.

We propose to learn a global max-margin classifier f that tells the positive class from the negative class while respecting the classification uncertainty u_i of each exemplar. Solving the dual formulation, we show that Power SVM seeks the closest points between reduced convex hulls of positive and negative exemplars in the feature space augmented by the classification uncertainty.

2.1. Primal: Parallel Planes of Max Separation

We first introduce notations. Let $'$ denote matrix transposition. We have two sets of features and their uncertainties:

$$\begin{aligned}
&\text{positive feature set:} && A_{d\times m} = [x_1, \ldots, x_m],\\
&\text{negative feature set:} && B_{d\times n} = [x_{m+1}, \ldots, x_{m+n}],\\
&A\text{'s uncertainty:} && U_{m\times 1} = [u_1, \ldots, u_m]',\\
&B\text{'s uncertainty:} && V_{n\times 1} = [u_{m+1}, \ldots, u_{m+n}]'.
\end{aligned}$$

We represent a linear classifier by two parallel bounding planes parametrized by (w, a, b), where $w_{d\times 1}$ is the normal of the planes, and a and b are the offset values for the positive and negative planes (Fig.2). We would like all the positive features A to lie on the positive side of the positive plane a, all the negative features B to lie on the negative side of the negative plane b, and the two planes to be maximally separated from each other. Since the distance between two planes is $\frac{a-b}{\|w\|}$, we maximize $a - b$ and minimize $\|w\|$ simultaneously. We introduce slack variables $p_{m\times 1}$ and $q_{n\times 1}$ to allow exemplars to cross over to the other sides of their bounding planes, if A and B are linearly inseparable. Such cross-overs should be minimized.

We formulate uncertainty u_i as the amount of allowance for $f(x_i)$ to reach the bounding plane of its class: The larger the classification uncertainty, the more tolerance for $f(x_i)$ to be located outside the bounding plane.

Taking the above considerations together with constant D weighing the importance of cross-overs, we propose the following Power SVM classifier that maximally separates two parallel bounding planes:

$$\begin{aligned}
\min_{w,a,b,p,q} \quad & \varepsilon = \tfrac{1}{2}w'w - (a-b) + D(p'1_m + q'1_n)\\
\text{s. t.} \quad & A'w + U \geq a1_m - p\\
& B'w - V \leq b1_n + q\\
& p \geq 0_m, \quad q \geq 0_n,
\end{aligned} \tag{1}$$

where a constant with subscript m (e.g. 1_m) denotes the $m \times 1$ vector of the same constant. Note that Power SVM becomes the classical SVM when $U = 0_m, V = 0_n$, where the decision boundary defined by two parallel planes of maximal separation (w, a, b) is equivalent to the one defined by a single plane with normal w and threshold $\frac{a+b}{2}$ [3].

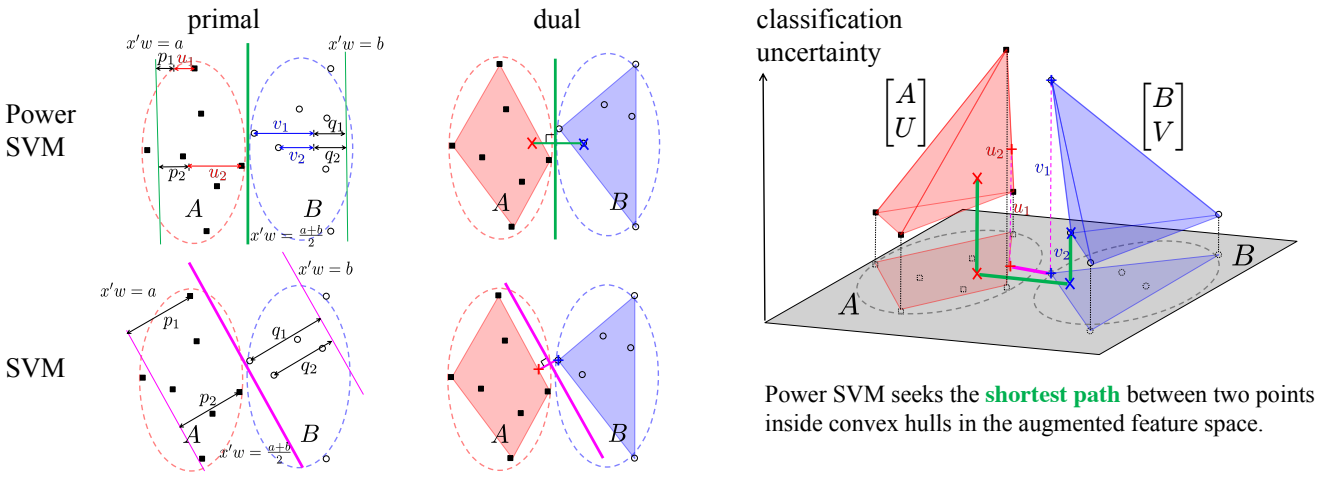

Figure 2. Power SVM (green, with u_i) and SVM (magenta, without u_i) on separating 2D point clouds (filled squares vs. open circles) sampled from Gaussian distributions (dashed ellipses). The goal of the primal is to find two parallel bounding planes of max separation between point clouds, with Power SVM additionally giving points different allowances to reach their bounding planes. The goal of the dual is to find the closest points between two convex hulls (shaded polygons), which lie in the 2D feature space for SVM and lie in the 3D augmented feature space for Power SVM. For the latter, the path length consists of vertical elevations of the two 3D points (along the 3rd dimension of classification uncertainty) and half the squared distance between the points' projection onto the horizontal feature plane. With the two extra uncertainty costs, the SVM optimum (magenta) is no longer optimal for Power SVM.

We call this optimization problem *Power SVM*, since each x_i is no longer a singular point in the feature space as in SVM, but a super-point with distinctive constraining power in the classifier space, just as points in a Voronoi Diagram become balls of varying radii in a Power Diagram.

2.2. Dual: Shortest Path between Convex Hulls

We solve Power SVM not in the classifier's parameter space,but in the exemplar space with $m+n$ dual variables (α, β). The dual formulation leads to a clear geometrical interpretation of finding the shortest path between reduced convex hulls of positive and negative exemplars, in the feature space augmented by the classification uncertainty.

The Power SVM primal has the following Lagrangian, with non-negative multipliers $\alpha_{m\times 1}$, $\beta_{n\times 1}$, $\xi_{m\times 1}$, $\eta_{n\times 1}$:

$$\begin{aligned}\max_{\alpha,\beta,\xi,\eta} \quad \min_{w,a,b,p,q} \; & L(w,a,b,p,q,\alpha,\beta,\xi,\eta) \\ = \; & \frac{1}{2}w'w - (a-b) + D(p'1_m + q'1_n) \\ & - \alpha'(A'w + U - a1_m + p) \\ & - \beta'(-B'w + V + b1_n + q) - \xi'p - \eta'q. \quad (2)\end{aligned}$$

Setting L's derivatives to 0, we have for the optimum:

$$L_w = w - A\alpha + B\beta = 0 \quad \Rightarrow w = A\alpha - B\beta \quad (3)$$
$$L_a = -1 + \alpha'1_m = 0 \quad \Rightarrow \alpha'1_m = 1 \quad (4)$$
$$L_b = 1 - \beta'1_n = 0 \quad \Rightarrow \beta'1_n = 1 \quad (5)$$
$$L_p = D1_m - \alpha - \xi = 0 \quad \Rightarrow \xi = D1_m - \alpha \quad (6)$$
$$L_q = D1_n - \beta - \eta = 0 \quad \Rightarrow \eta = D1_n - \beta \quad (7)$$

The last two equations show that $\alpha \le D1_m$ and $\beta \le D1_n$. Eliminating w, a, b, p, q, ξ, η, we reach the dual form:

$$\begin{aligned}\min_{\alpha,\beta} \quad & -L = \frac{1}{2}\|A\alpha - B\beta\|^2 + (U'\alpha + V'\beta) \\ \text{s.t.} \quad & \alpha'1_m = 1, \quad \beta'1_n = 1, \\ & 0_m \le \alpha \le D1_m, \quad 0_n \le \beta \le D1_n.\end{aligned} \quad (8)$$

α and β can be interpreted as combination coefficients, thus $A\alpha$ and $B\beta$ represent points in the convex hulls of positive and negative exemplars. D is an upper bound on α, β: Any $D > 1$ is equivalent to $D = 1$; when $D < 1$, we have reduced convex hulls; when $D < \frac{1}{\min(m,n)}$, there is no feasible solution. The effective range of D is thus $[\frac{1}{\min(m,n)}, 1]$. At the lowest extreme $D = \frac{1}{\min(m,n)}$, the convex hull of positive exemplars (when $m \le n$) or negative exemplars (when $m > n$) is reduced to a single centroid point.

The objective $-L$ can be interpreted as the length of the path connecting two $(d+1)$-dimensional points in the d-dimensional feature space augmented by an additional dimension for classification uncertainty. The two points lie inside the reduced convex hulls of positive and negative exemplars $\left(\begin{bmatrix} A \\ U \end{bmatrix}$ and $\begin{bmatrix} B \\ V \end{bmatrix}\right)$ respectively. The path between the two points consists of three segments, two along the uncertainty dimension with length $U'\alpha + V'\beta$ and one inside the original feature space with length $\frac{1}{2}\|A\alpha - B\beta\|^2$).

Since the classifier $w = A\alpha - B\beta$ is a linear combination of exemplars, Power SVM can be kernelized just like SVM.

3. Exemplar Classification Uncertainty

The classification uncertainty u_i indicates how easily the exemplar with feature x_i might be confused with those from the opposite class. It is different from data uncertainty [4] that captures measurement noise in x_i or data fuzziness [18] that captures labeling noise in y_i. Both measure the data quality of each exemplar itself, whereas our classification uncertainty is concerned with the quality of one exemplar against other exemplars and captures its discrimination capacity in the classifier's output space.

The key here is that, we can obtain an informative estimate of how discriminative an exemplar is without actually knowing the desired classifier in advance. In the human visual experience, given a few frontal views of persons of interest, we pretty much know before seeing other photos of variations that those with distinctive looks will be identified more readily and confidently.

Among many possible alternatives, we propose an approach based on local classification performance for individual exemplars. Without compromise among exemplars, such local discrimination performance provides a lower bound on the global classification uncertainty, which can then be used to shape the desired global classifier.

Given m positive and n negative exemplars, we train m local SVM classifiers that each maximally separates a positive exemplar from all the n negative exemplars (Fig.3). We normalize each classifier so that all the f_i responses are comparable across i's. For example, if f_i is a linear SVM parametrized by normal w_i and threshold t_i, we scale $f_i = (w_i, t_i)$ so that w_i has a norm of 1.

The classification uncertainty for positive exemplar x_i is

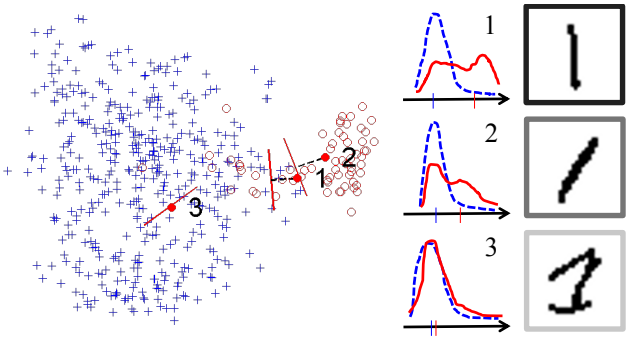

Figure 3. Uncertainty U for positive exemplars. The scatter plot is a 2D PCA visualization of 784-dimensional positive exemplars (MNIST digit 1, ○), negative exemplars (other digits, +), and local classifier f_i's on three marked exemplars (connected to their solid-line decision boundaries). The two curves are the distributions of f_i over the positive (red solid line) and negative (blue dashed line) exemplars, and their average separation is indicated by the gap between two ticks on the x-axis. The smaller the separation, the larger the uncertainty for the exemplar (lighter image outline).

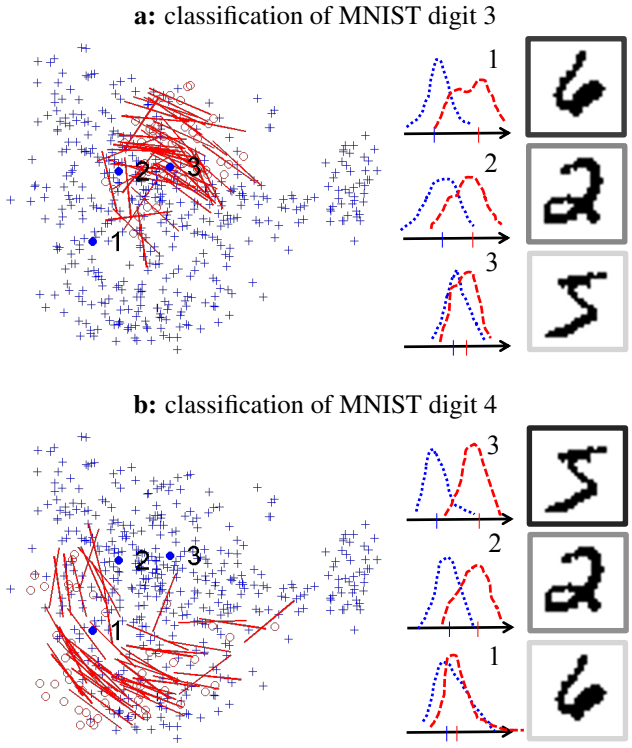

Figure 4. Uncertainty V for negative exemplars. Same convention as Fig.3, except that the two curves are the distributions of all f_i's over the positive exemplars (red dashed line) and over the negative exemplar itself (blue dotted line). An exemplar could assume different uncertainty values in different classification tasks.

characterized by how well its local classifier f_i separates the m positive exemplars from the n negative exemplars (Fig.3). Let s_i be the difference between the average positive response and the average negative response for f_i:

$$s_i = \sum_{t=1}^{m} \frac{f_i(x_t)}{m} - \sum_{t=m+1}^{m+n} \frac{f_i(x_t)}{n}, \qquad i \le m. \quad (9)$$

A large s_i means that, while f_i is designed just to separate a single positive exemplar x_i from negative exemplars, it in fact well separates the entire positive class from the negative class. The exemplar x_i is thus rather representative of the positive class and has a large discrimination capacity.

Likewise, the classification uncertainty for negative exemplar x_j is characterized by how well this exemplar can be separated from the positive class by all the local classifiers. Let s_j be the average difference over all f_i's between each classifier f_i's response on the positive class and on x_j:

$$s_j = \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=1}^{m} \frac{f_i(x_t)}{m} - f_i(x_j) \right), \qquad j > m. \quad (10)$$

A large s_j means that, all f_i's tend to put x_j in the negative class. The exemplar x_j is thus rather representative of the negative class and has a large discrimination capacity.

For either positive exemplar x_i or negative exemplar x_j, its classification uncertainty u_i or u_j is inversely correlated with the separation s_i or s_j, and can be defined as:

$$u_i \propto \max_{1 \le t \le m} s_t - s_i, \qquad i \le m; \tag{11}$$

$$u_j \propto \max_{m+1 \le t \le m+n} s_t - s_j, \qquad j > m. \tag{12}$$

Since u_i and u_j often have different value distributions, we normalize them separately so that each has a range of [0,1].

For k-way classification, we learn k one-vs-all global classifiers. If each class has N exemplars, we first learn Nk local classifiers, each separating an exemplar from the rest $N(k-1)$ exemplars. For each of the k global classifier, we then derive the classification uncertainty for N positive exemplars and $N(k-1)$ negative exemplars, and apply Power SVM to obtain the global classifier. Each exemplar thus has a total of k classification uncertainty measurements, one as a positive exemplar in its own class, and $k-1$ as a negative exemplars in the rest $k-1$ classes. Fig.4 shows that the same exemplars can assume very different uncertainty in different one-vs-all classification tasks.

The running time of our Power SVM is dominated by training Nk (especially nonlinear) local classifiers. However, it could be done off-line and more efficiently with a linear approximation method [21]. In addition, solving a local classifier for a single positive exemplar is much easier and faster than for a general category of exemplars. We are investigating simpler uncertainty estimation methods.

4. Experimental Results

We solve Power SVM in its dual form, and implement it using libsvm library [6]. We conduct a series of experiments investigating how Power SVM performs with its own parameter choice, with various forms of exemplar classification uncertainty input, and in comparison with SVM and uncertainty-weighted SVM on large-scale object recognition and scene categorization tasks.

4.1. Power SVM Parameter Choice

Power SVM has one parameter D. In the primal, D weighs the importance of misclassification over class separation, whereas in the dual, D controls the extent of the reduced convex hulls. As D increases within its effective range of $[\frac{1}{\min(m,n)}, 1]$, Power SVM searches the shortest path between two increasingly larger convex hulls.

We study the effect of D on the MNIST digit dataset. We train a linear classifier in the feature space of 784 concatenated pixel values with L_2 normalization. We use 200 out of $60,000$ training images and evaluate the mean classification error on the training set as well as $10,000$ test images.

The digit recognition rate improves monotonically and then drops to a plateau as D increases (Fig.5a). This result can be understood from a geometrical point of view:

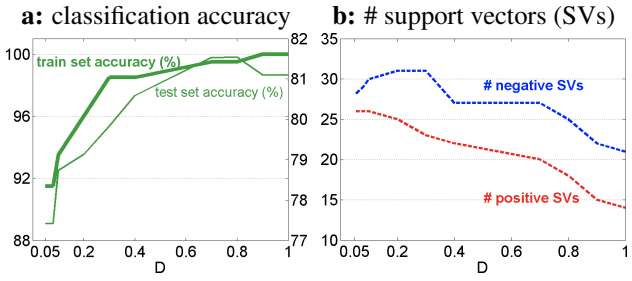

Figure 5. Power SVM parameter D controls the importance of misclassification (in the primal) and the extent of reduced convex hulls (in the dual). When D is at the lowest limit, the reduced convex hull of positive exemplars is just its centroid and all the exemplars are support vectors. As D increases, the reduced convex hulls expand and more feasible solutions are considered, the classification accuracy increases on the training and test sets. As the optimum is found more on convex hull vertices and with fewer exemplars, overfitting is likely to occur, and the performance drops.

The optimum is found in an increasingly larger feasible region, thus the classification performance keeps improving initially; as the reduced convex hulls expand towards the fullest, the optimum moves towards convex hull vertices (training data) and overfitting over fewer exemplars (Fig.5b) is more likely to reduce the test performance. In the rest of experiments, we use cross-validation to find the optimal D.

4.2. Various Classification Uncertainty Comparison

Power SVM relies on good estimates of exemplar classification uncertainty. We compare our uncertainty method with three other simple approaches: 1) Uniform uncertainty. Power SVM in this case is reduced to the regular SVM, where every exemplar assumes 0 uncertainty. 2) Random uncertainty. This baseline case helps establish the utility of uncertainty. 3) Local-frequency uncertainty. This case simply measures the proportion of nearest neighbours in the same category: If all the neighbours are of the same category as the exemplar, its classification uncertainty is 0.

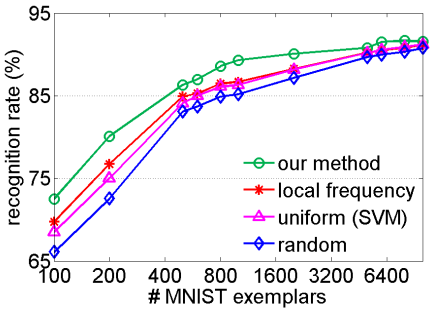

Figure 6. Our exemplar classification uncertainty outperforms local frequency uncertainty (percentage of exemplars of a different category in 10 neighbours), uniform uncertainty (0 for every exemplar), and random uncertainty (uniformly distributed over [0,1]). The benefits are consistent but diminishing with more exemplars.

Our uncertainty outperforms the local frequency uncertainty, which only captures the categorical exemplar density but does outperform uniform uncertainty assumed by SVM (Fig.6). The random uncertainty is not only uninformative but also damaging, since its performance is worse than the uniform uncertainty of the regular SVM.

With an increasing number of exemplars, the recognition rate always improves, and the benefits of good classification uncertainty diminish. By acknowledging the distinction in the classification uncertainty among exemplars, Power SVM in fact maximizes the utility of each exemplar and is particularly effective when the training size is small.

We also compare our uncertainty with human classification uncertainty on various features. If an image can be categorized by human subjects accurately, the classification uncertainty is inherently low. While we do not know which feature the human vision uses, the human classification accuracy nevertheless provides a lower bound on the exemplar classification uncertainty in the feature we investigate. For a binary classification task, if the human accuracy is a, we derive exemplar uncertainty as $2(1-a)$, with $a = 0.5$ (random guessing) mapped to the maximal uncertainty of 1.

We train Power SVM on the 50 indoors vs. 50 outdoors dataset obtained from [25], and test it on 1,000 images sampled from 14 of the 15 scene categories dataset in [17], with the *store* category excluded due to its indoor/outdoor ambiguity. We provide indoor/outdoor class labels based on the semantic meaning of the category name.

Consistent with Fig.6, Table 1 shows that our uncertainty outperforms all the other data-driven uncertainty. It also outperforms human classification uncertainty on GIST and HOG features. There are two contributing factors: 1) Unlike human uncertainty, our uncertainty is specifically tuned to the feature used for final classification; 2) As a lower bound of the classification uncertainty, human uncertainty saturates at 0 for about 25% of exemplars, whereas our data-driven uncertainty has a much finer and richer separation between training data, promoting a larger selection variety of support vectors for the final classifier.

uncertainty type	**GIST**	**Sparse SIFT**	**HOG**
our method (%)	**81.9**	83.4	**89.0**
human (%)	80.2	**84.5**	87.1
local frequency (%)	81.0	83.3	88.0
uniform (SVM)(%)	77.6	82.9	87.3
random (%)	75.2	81.2	85.2

Table 1. Average test accuracy of Power SVM with five types of exemplar classification uncertainty (rows) on an indoor-outdoor scene categorization task on three types of feature-classifier settings (columns): GIST with RBF kernel, sparse SIFT with intersection kernel, and HOG with intersection kernel.

Human uncertainty on sparse SIFT delivers better performance. It is likely a more accurate estimate, since the human performance data is collected in an ultra-rapid (16-millisecond viewing time) categorization task, i.e. only sparse and salient features are best processed with the human visual system. For dense features such as GIST and HOG, our data-driven uncertainty is more relevant.

In short, it is possible to estimate good exemplar classification uncertainty from data or an external source such as human classification accuracy. Using it as a constraint on the desired global classifier allows Power SVM to deliver much improved results with limited training data.

4.3. Power SVM vs. SVM and Weighted SVM

SVM is a special case of Power SVM, where all the exemplars have uniformly zero classification uncertainty. We gain further understanding into Power SVM by comparing its results with SVM's on 10-class MNIST digit recognition in a simple setting which trains linear classifiers on concatenated pixel values over a total of 200 exemplars.

Our Power SVM improves accuracy over SVM for all 10 MNIST digits, with the largest gain (+12.4%) for digit 9 and smallest gain (+1.2%) for digit 1 (Fig.7a). We observe

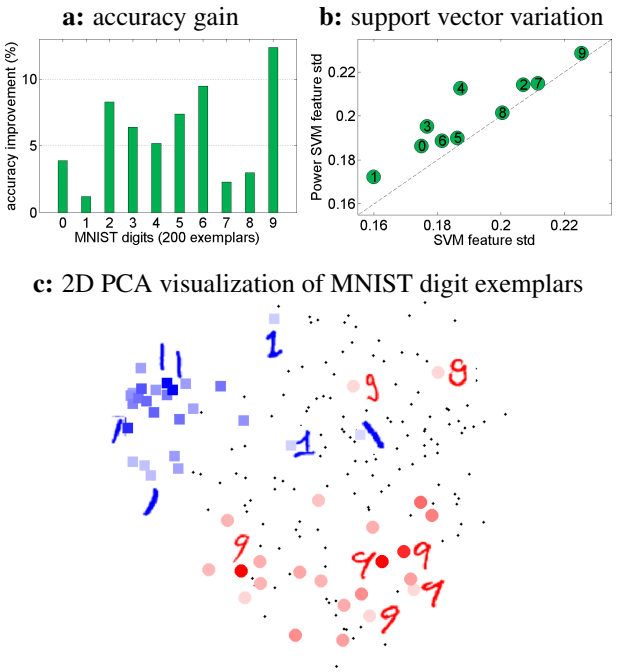

Figure 7. Power SVM with our uncertainty shows consistent improvement over SVM on all 10 MNIST digits (**a**). The number of exemplars is chosen according to Fig.6. **b)** We compare the mean standard deviation of support vectors for positive exemplars between Power SVM and SVM on top 10 PCA dimensions, and the former is larger for all 10 digits. **c)** We visualize the distribution of exemplars as in Fig.3, with digits 1 (blue squares) and 9 (red dots) highlighted according to the uncertainty. A few most certain (i.e. reddest, bluest) and uncertain ones are labeled with their images.

that Power SVM tends to select a wider range of exemplars as support vectors (Fig.7b), and the larger the uncertainty variation and exemplar distribution (Fig.7c), the larger the accuracy gain for Power SVM over SVM.

We compare Power SVM with an SVM variant where the uncertainty is treated as per exemplar weight in the cost function (the larger the uncertainty, the smaller the weight):

$$\begin{array}{ll} \min\limits_{w,t,p,q} & \varepsilon = \frac{1}{2}w'w + C(p'(1_m - U) + q'(1_n - V)) \\ \text{s. t.} & A'w - t \geq \quad 1_m - p, \quad p \geq 0_m, \\ & B'w - t \leq -1_n + q, \quad q \geq 0_n. \end{array} \tag{13}$$

This weighted SVM is similar to [18] and [14], developed to address class label noise and uneven class sizes respectively.

It is not only always better to utilize an informative exemplar classification uncertainty in SVM, but Power SVM is also more effective than weighted SVM at utilizing the same uncertainty (Fig.8a). This result can be justified theoretically. In the primal, weighted SVM can be viewed as duplicating higher certainty exemplars with weak impact on support vectors, whereas Power SVM directly pushes higher certainty exemplars away from the decision boundary. In the dual, weighted SVM changes the shape of convex hulls (often slightly) while Power SVM modifies the distance measure for the shortest path between convex hulls.

Power SVM consistently outperforms weighted SVM and SVM in two scene classification tasks with kernel SVM

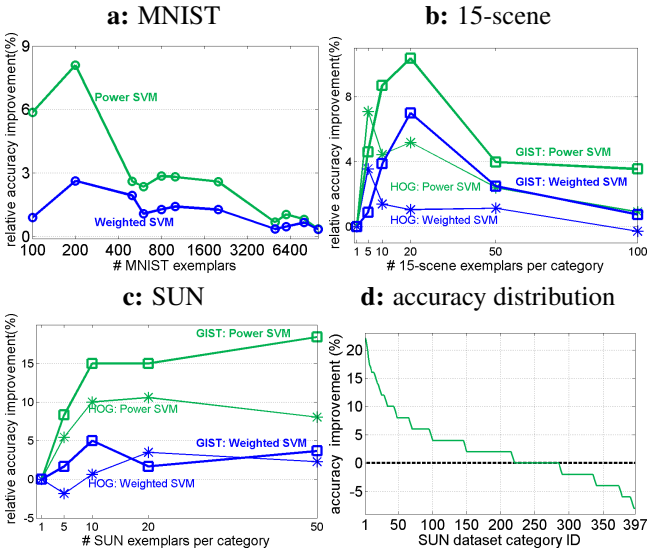

Figure 8. Power SVM (green) consistently outperforms weighted SVM (blue), both better than SVM (at 0% level as a baseline). The relative accuracy improvement is calculated as gain over the SVM baseline. **a)** For the 10,000 MNIST test images, the SVM accuracy increases from 5.9% to 0.3%. **b,c)** For the 1500 15-scene and 19850 SUN test images on both GIST (squares on thick lines) and HOG (stars on thin lines) features, the SVM accuracy increases from 1% to 10.3% and from 5.4% to 18.4% respectively. At 1 exemplar per category, the three SVM methods become the same. **d)** Power SVM improves over kernel SVM on most SUN categories.

[28] on GIST and HOG features (Fig.8b,c). The largest relative improvement of Power SVM over SVM (normalized by the performance of SVM) is 10.3% for the 15-scene dataset [17] with GIST at 20 exemplars per category, and 18.4% for the SUN dataset [28] with HOG at 50 exemplar per category. The latter improvement is substantial for most categories, with the largest absolute gain at 22% and loss at 8% among all the 397 categories (Fig.8d).

Fig.9 shows sample results on categories with the largest accuracy gain and loss. Indoor scenes with larger intra-category variations get bigger improvement compared to outdoor nature scene categories with smaller variation. Positive exemplars of larger uncertainty often have extreme lightening and smaller fields of view.

We compare our Power SVM to an entirely exemplar-based approach [22]: Exemplar-centric classifiers are first learned and calibrated so that their scores on a test image can be compared and the winning exemplar's label is the global classification result. Its accuracy on the SUN dataset is only 12.3% compared to our 28.3% on HOG feature using 50 training exemplars per category. This result suggests that, while uncertainty derived from local classifiers helps, integrating local discriminative information into one global discrimination framework provides greater benefits.

It is theoretically and empirically compelling that Power SVM, a simple idea on SVM using estimated exemplar classification uncertainty, can deliver an efficient global classifier that generalizes most effectively from a few exemplars.

Conclusions. We present a classification framework that computes exemplar classification uncertainty based on its local discrimination against other categories, and then uses it as constraints for learning a desirable global classifier. We propose Power SVM which maximizes the separation between parallel bounding planes on two classes of exemplars and can be solved by finding the shortest path between convex hulls in the feature space augmented by the dimension of classification uncertainty. We demonstrate that Power SVM outperforms SVMs on multiple categorization tasks, especially when exemplars have a wider range of local discriminability and when the training data size is small.

Acknowledgements. This research was funded by NSF CAREER IIS-0644204 and by Boston College to Stella Yu. Part of Teng's work was supported by NSF CCF 1111270.

References

[1] S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. In *NIPS*, 2010.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[3] K. P. Bennett and E. J. Bredensteiner. Duality and geometry in svm classifiers. In *ICML*, 2000.

[4] J. Bi and T. Zhang. Support vector classification with input data uncertainty. In *NIPS*, 2004.

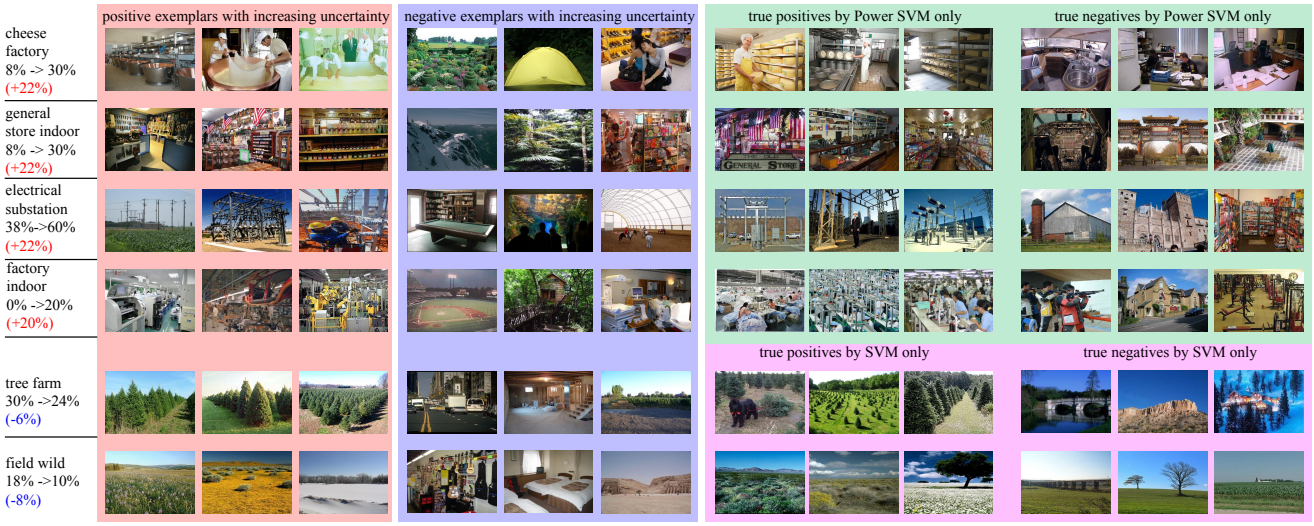

Figure 9. Sample results on the SUN dataset. Top four and bottom two rows show categories in which accuracies increase and decrease the most. Column 1 shows the change in accuracy from SVM to Power SVM. Columns 2-3 show sample positive and negative exemplars with increasing uncertainty. Columns 4-5 show test image samples in different classification scenarios.

[5] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *ECCV*, 2008.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[7] S.-W. Cheng, T. K. Dey, H. Edelsbrunner, M. A. Facello, and S.-H. Teng. Silver exudation. *J. ACM*, 47(5):883–904, Sept. 2000.

[8] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

[9] C. Domeniconi and D. Gunopulos. Adaptive nearest neighbor classification using support vector machines. In *NIPS*, 2001.

[10] L. F. Fei, R. Fergus, and P. Perona. One-Shot Learning of Object Categories. *PAMI*, 28(4):594–611, 2006.

[11] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*, 2009.

[12] T. Gao and D. Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *ICCV*, 2011.

[13] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Trans. Graph.*, 26(3):4, 2007.

[14] Y. Huang and S. Du. Weighted support vector machine for classification with uneven training class sizes. In *ICMLC*, pages 5245–8, 2005.

[15] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, 2007.

[16] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*. 2010.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[18] C.-F. Lin and S.-D. Wang. Fuzzy support vector machines. *ITNN*, 13(2):464–471, 2002.

[19] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009.

[20] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*, 2008.

[21] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[22] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[23] M. Marszalek and C. Schmid. Constructing category hierarchies for visual recognition. In *ECCV*, 2008.

[24] E. G. Miller, N. E. Matsakis, and P. A. Viola. Learning from one example through shared densities on transforms. In *CVPR*, 2000.

[25] C. Pavlopoulou and S. X. Yu. Classification and feature selection with human performance data. In *ICIP*, 2010.

[26] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 30(11):1958–1970, 2008.

[27] S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, 2010.

[28] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[29] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.

[30] W. Zhang, P. Srinivasan, and J. Shi. Discriminative image warping with attribute flow. In *CVPR*, 2011.

[31] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*. 2004.