Practical Analysis of the Universum SVM Learning

Vladimir Cherkassky
University of Minnesota
Minneapolis, MN 55455
cherk001@umn.edu

Sauptik Dhar
University of Minnesota
Minneapolis, MN 55455
dharx007@umn.edu

The idea of 'inference through contradictions' was introduced by Vapnik[1] in order to incorporate a priori knowledge into the learning process. This knowledge is introduced via additional unlabeled data samples (called virtual examples or the Universum) that are used along with labeled training samples, to perform an inductive inference. For example, if the goal of learning is to discriminate between handwritten digits 5 and 8, one can introduce additional 'knowledge' in the form of other handwritten digits 0, 1, 2, 3, 4, 6, 7, 9. Obviously, such Universum samples contain certain information about handwritten digits, but they do not have the same distribution as labeled training samples. The idea that 'a good universum needs to be positioned *in-between* the two classes' is implicit in Vapnik's original formulation [2] and in the loss function which penalizes Universum samples that are close to either class. However, after introduction of U-SVM [2, 3] several researchers tried to quantify this notion explicitly, in terms of analytic properties of the Universum and labeled data. Sinz et al [4] showed that the optimal decision boundary of the U-SVM tends to make the normal vector orthogonal to the principal direction of the Universum data set. This condition holds for both the original Vapnik's Universum SVM (U-SVM) formulation and for the least-squares U-SVM, where the squared loss function is used for labeled and Universum samples.

Our work pursues the same general objective as [4], i.e. the characterization of a good Universum for Vapnik's original formulation. However, we take a more practical and specific approach. That is, we ask the following questions:

i. Can a given Universum data set improve generalization performance of a standard SVM classifier trained using only labeled data?
ii. Can we provide practical conditions for (i), based on the properties of the Universum data and labeled data?

This approach is more suitable for non-expert users, because:

- practitioners are interested in using U-SVM only if it provides an improvement over standard SVM;
- the problem of (full-blown) model selection for the U-SVM is alleviated, because its two parameters (kernel and regularization parameter C) are tuned separately, during training standard SVM classifier.

Proposed strategy for analyzing practical conditions for the effectiveness of the U-SVM is outlined below:

a. estimate standard SVM classifier for a given (labeled) training data set. Note that this step includes optimal model selection, i.e. optimal tuning of the regularization parameter C and kernel;
b. generate low-dimensional representation of training data [5] by projecting it onto the normal direction vector of the SVM decision boundary estimated in (a);
c. project the Universum data onto the normal direction vector (of SVM decision boundary), and analyze projected Universum data in relation to projected training data.

Then statistical properties of the projected Universum data, relative to labeled training data, may suggest whether using this Universum will provide improvement vs. standard SVM estimated in step (a). In particular, a Universum data set will be effective if its histogram of projections satisfies two conditions:

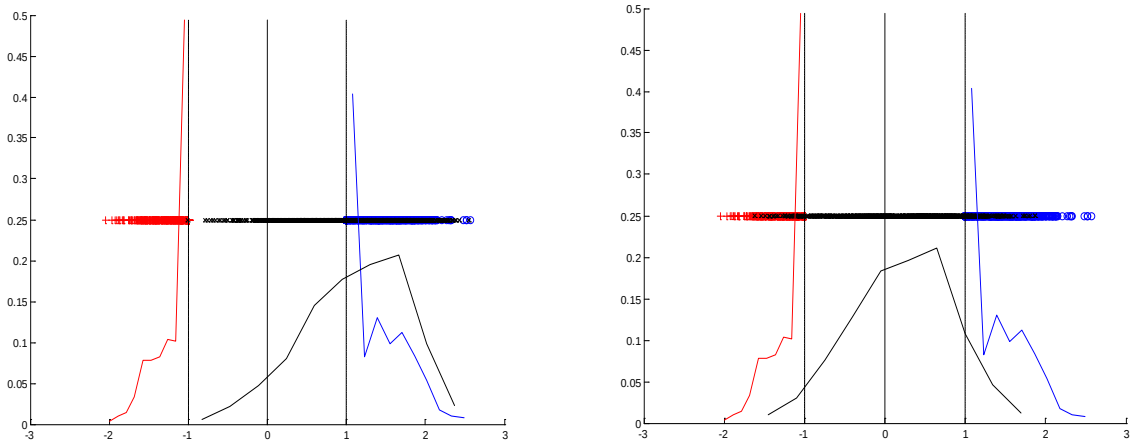
(C1) It is symmetric relative to the (standard) SVM decision boundary, and

(C2) It has wide distribution between margin borders denoted as points -1/+1 in the projection space.

We emphasize that the effectiveness of Universa can be evaluated only in the context of particular labeled training data. For example, consider classification of handwritten digits '5' and '8' using the MNIST data. The goal is to investigate the effectiveness of two types of Universa: handwritten digits 1 and 3, and to explain their effectiveness by analyzing histograms of projections of both labeled and Universum data sets. This set of experiments used training/validation set size of 1,000 (500 per class), Universum set size 1,000, and test set size 1,866. Model selection for standard RBF SVM classifier and for U-SVM is performed using the validation data set. Each experiment is repeated 10 times with different random realizations of training/validation/Universum samples, and the average test error (and its standard deviation in parenthesis) is reported. The test error rates of SVM and U-SVM are shown in Table 1, and the typical histograms of projections for training data and Universum data are shown in Fig. 1.

TABLE 1. TEST ERROR RATES FOR MNIST DATA WITH DIFFERENT UNIVERSA. TRAINING SET SIZE IS 1,000 SAMPLES.

	SVM	U-SVM (digit 1)	U-SVM (digit 3)
Test error	1.47% (0.32%)	1.31% (0.31%)	1.01% (0.28%)



(a) digit 1 (b) digit 3

Fig. 1. Univariate histogram of projections for 2 different types of Universa. *Legend:* histogram of projections of digit '5' samples is shown in red, digit '8' samples - in blue, Universum samples – in black. SVM margin borders are marked as -1/+1.

The histograms of projections shown in Fig. 1 suggest that digit 1 Universum is less effective than digit 3 because its projections are more biased towards the distribution of digit 8. The Universum samples for digit 3 is more widely and symmetrically distributed inside the margin borders, so it is expected to provide better performance (than digit 1 Universum). These findings are consistent with the empirical results in Table 1, showing no statistically meaningful improvement for digit 1 Universum, and a significant improvement for digit 3.

Next, we provide theoretical justification of our practical conditions for the effectiveness of Universum learning. Following [2,3], the linear U-SVM formulation can be given as (setting $\varepsilon = 0$ for simplicity).

$$\min. \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} H(y_i f(\mathbf{x}_i)) + C^*\sum_{j=1}^{m}\left| f(\mathbf{x}_j)\right| \qquad \begin{array}{l} i = 1...n \text{ (Total \# of Training samples)} \\ j = 1...m \text{ (Total \# of Universum samples)} \\ \text{H(t)=max\{0,1-t\} is the Hinge Loss} \end{array} \quad \ldots\ldots (1)$$

This optimization formulation (1) can be approximated as,

$$\equiv \min. \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} H(y_i f(\mathbf{x}_i)) + C^*(m\sigma_U^2 + m^2\overline{u}^2)^{\frac{1}{2}} \qquad \ldots\ldots (2)$$

where, $\overline{u} = \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{x}_j)$ and $\sigma_U^2 = \frac{1}{m}\sum_{j=1}^{m}(f(\mathbf{x}_j) - \overline{u})^2$ are the mean and the variance of the projections of the universum samples onto the normal weight vector (**w**). Typically, for standard SVM models estimated from High Dimensional Low Sample Size data $\sum_{i=1}^{n} H(y_i f(\mathbf{x}_i)) \approx 0$. So U-SVM additionally tries to find a direction for which the mean and the variance of the projected values for the universum samples are small. Our conditions specify that for a case when the projection of the Universum data is symmetric relative to the (standard) SVM decision boundary (i.e $\overline{u} \approx 0$). In addition, U-SVM can provide a possible improvement over SVM by minimizing the wide distribution of the universum samples between SVM margin borders , i.e. by minimizing large σ_U^2 in (2).

REFERENCES

[1] V. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.
[2] V. Vapnik, Estimation of Dependencies Based on Empirical Data. Empirical Inference Science: Afterword of 2006. NY: Springer, 2006.
[3] J. Weston, R. Collobert, F. Sinz, L. Bottou, and V. Vapnik, "Inference with the Universum," Proc. ICML, 2006, pp. 1009-1016.
[4] F. Sinz, O. Chapelle, A. Agarwal, and B. Schölkopf, "An Analysis of Inference with the Universum," In Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems(NIPS), pp 1–8, 2008.
[5] V. Cherkassky, and S. Dhar "Simple Method for Interpretation of High-Dimensional Nonlinear SVM Classification Models," Proc. DMIN, July 2010, pp. 267-272.

Topic: Learning algorithms
Preference: Oral