Cost-Sensitive Universum-SVM

Sauptik Dhar
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis MN 55455 USA
e-mail: dharx007@umn.edu

Vladimir Cherkassky, Fellow, IEEE.
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis MN 55455 USA
e-mail: cherk001@umn.edu

***Abstract*— Many applications of machine learning involve analysis of sparse high-dimensional data, where the number of input features is larger than the number of data samples. Standard classification methods may not be sufficient for such data, and this provides motivation for non-standard learning settings. One such new learning methodology is called Learning through Contradictions or Universum support vector machine (U-SVM) [1, 2]. Recent studies [2-10] have shown U-SVM to be quite effective for such sparse high-dimensional data settings. However, these studies use balanced data sets with equal misclassification costs. This paper extends the U-SVM for problems with different misclassification costs, and presents practical conditions for the effectiveness of the cost sensitive U-SVM. Finally, several empirical comparisons are presented to illustrate the utility of the proposed approach.**

Keywords- Cost-sensitive SVM, learning through contradiction, support vector machines, Universum SVM.

I. Introduction

Many modern machine learning applications involve predictive modeling of high-dimensional data, where the number of input features exceeds the number of data samples used for model estimation. Such high-dimensional data sets present new challenges for classification methods.

Recent studies have shown Universum learning to be particularly effective for such high-dimensional low sample size data settings [2-10]. Most of these studies use balanced data sets with equal misclassification costs. That is, the number of positive and negative labeled samples is (approximately) the same, and the relative cost of false positive and false negative errors is assumed to be the same. However, many practical applications involve unbalanced data and different misclassification costs, i.e., credit card fraud detection, intrusion detection, oil-spill detection, disease diagnosis etc [11-13]. So there is a need to extend the Universum learning to handle such settings.

Researchers have introduced many techniques to deal with problems involving unequal misclassification costs and unbalanced data scenarios [11-13]. Typically, most of these methods follow two basic frameworks,

- *Cost-Sensitive Learning*: where the costs of misclassification and the ratio of unbalance in the data are introduced directly into the learning algorithm.
- *Sampling based Approach (Oversampling/ Undersampling)*: where the samples of a particular class are replicated until the training set has equal number of positive and negative samples or an equivalent misclassification cost ratio [13].

In this paper we follow the direct approach of introducing the cost-ratios into U-SVM formulation. We extend Vapnik's original formulation for U-SVM [1, 2] to include different misclassification costs. Further, we provide characterization of a good Universum for the proposed cost-sensitive U-SVM formulation. Our approach follows a practical strategy that tries to address the following questions:

i. Can a given Universum data set improve generalization performance of the cost-sensitive SVM classifier [14, 15] trained using only labeled data?
ii. Can we provide practical conditions for (i), based on the geometric properties of the Universum data and labeled training data?

This approach is more suitable for non-expert users, because practitioners are interested in using cost-sensitive U-SVM only if it provides an improvement over the existing cost-sensitive SVM. Our conditions for the effectiveness of cost-sensitive U-SVM extend the conditions for the effectiveness of the standard U-SVM introduced in [3].

The paper is organized as follows. Section II describes Vapnik's original formulation for U-SVM [1] and presents practical conditions for its effectiveness [3]. Section III presents new cost-sensitive U-SVM formulation and the practical conditions for its effectiveness. Section IV provides empirical results to illustrate these conditions, using both synthetic and real-life data sets. Finally, conclusions are presented in Section V.

II. Practical Conditions for standard U-SVM learning

The idea of Universum learning was introduced by Vapnik [1, 2] to incorporate a priori knowledge about admissible data samples. He introduced Universum learning for binary classification where in addition to labeled training data we are also given a set of unlabeled examples from the Universum. The Universum contains data that belongs to the same application domain as the training data. However, these samples are known not to belong to either class. These Universum samples are incorporated into inductive learning as explained next. Let us assume that labeled training data is linearly separable using large margin. Then the Universum samples can fall either inside the margin or outside the margin borders (see Fig. 1). Under U-SVM we favor

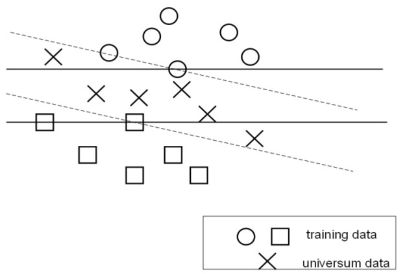

Fig.1. Two large-margin separating hyperplanes explain training data equally well, but have different number of contradictions on the Universum. The model with a larger number of contradictions should be selected.

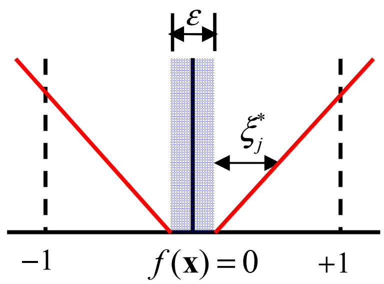

Fig. 2. The ε - insensitive loss for the Universum samples. Universum samples outside the ε -insensitive zone are linearly penalized using the slack variables ξ_j^*.

hyperplane models where the Universum samples lie inside the margin, as these samples do not belong to either class. Such Universum samples (inside the margin) are called contradictions, because they are falsified by the model (i.e., have non-zero slack variables for either class label).

The optimization formulation for U-SVM [1, 2], is presented in Box (1). For improved readability we show only linear parameterization; however it can be generalized to nonlinear case using kernels. Here, for labeled training data, we use standard SVM soft-margin loss with slack variables ξ_i. For the Universum samples $(\mathbf{x}_j^*)$, we penalize the real-valued outputs of our classifier using an ε – insensitive loss (shown in Fig. 2). Let ξ_j^* denote slack variables for samples from the Universum. Then the U-SVM formulation is given as,

$$\min_{\mathbf{w},b} R(\mathbf{w},b) = \frac{1}{2}(\mathbf{w}\cdot\mathbf{w}) + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{m}\xi_j^* \qquad \mathbf{(1)}$$

subject to constraints:

$$(\textit{training samples}):\quad y_i[(\mathbf{w}\cdot\mathbf{x}_i)+b] \geq 1-\xi_i$$

$$(\textit{universum samples}):\quad \left|(\mathbf{w}\cdot\mathbf{x}_j^*)+b\right| \leq \varepsilon + \xi_j^*$$

$$\xi_i \,, \geq 0, i = 1,\dots,n \qquad \xi_j^* \geq 0, j = 1,\dots,m$$

Here $\varepsilon \geq 0$ is user-defined. Parameters $C, C^* \geq 0$ control the trade-off between the margin size, the number of errors and the number of contradictions.

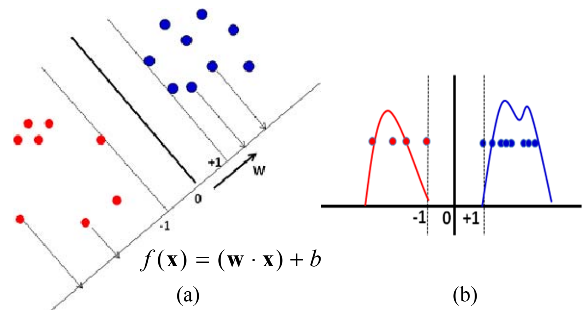

Fig. 3 (a) Projection of the training data shown in red and blue onto the normal weight vector ($\mathbf{w}$) of the SVM hyperplane. (b) Univariate histogram of projections. i.e. histogram of $f(\mathbf{x})$ values for training samples.

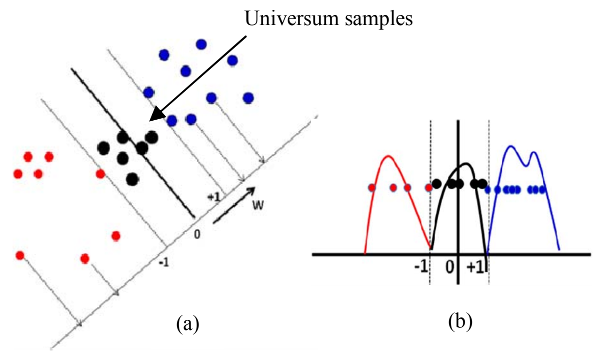

Fig. 4 (a) Projection of the universum data (shown in **black)** onto the normal weight vector ($\mathbf{w}$) of the SVM hyperplane. (b) Univariate histogram of projections of the universum samples (shown in **black**) along with the training samples (shown in red/blue).

TABLE 1. STRATEGY TO ANALYZE THE EFFECTIVENESS OF U-SVM [3]

1a.	estimate SVM classifier for a given (labeled) training data set. This step involves model selection for the C and kernel parameter.
1b.	generate low-dimensional representation of training data by projecting it onto the normal direction vector of SVM hyperplane estimated in (1a) (see Fig. 3).
1c.	project the Universum data onto the normal direction vector of the SVM hyperplane (see Fig. 4a).
1d.	analyze the histogram of projected Universum data in relation to projected training data (see Fig. 4b).

The solution to the optimization problem (1) defines the large margin hyper-plane $f(\mathbf{x}) = (\mathbf{w}^* \cdot \mathbf{x}) + b^*$ that incorporates a priori knowledge (i.e., Universum samples) into the final model. There are *two design factors* necessary for a successful practical application of U-SVM.

– *Model Selection*: which becomes rather difficult because the kernelized U-SVM has 4 tunable parameters: C , C^*, kernel parameter and ε (in contrast, standard SVM has only two tuning parameters).
– Alternatively, generalization performance of U-SVM may be affected by a bad choice of the Universum data.

In practice, it is difficult to separate these two factors. However, our recent paper [3] suggests a practical approach to this problem. The strategy for judging the effectiveness of a given Universum is based on analysis of the histogram of projections of the training and universum samples onto the normal direction of the SVM decision boundary (see Table 1). The benefits of this strategy (in Table 1) are two- fold. First, it simplifies the characterization of good Universum data. Specifically, based on the statistical properties of the projected Universum data relative to labeled training data (in step. 1d), we can formulate the conditions on whether using this Universum will improve the prediction accuracy of standard SVM estimated in step 1a. The practical conditions for the effectiveness of U-SVM [3] are provided in Table 2 and illustrated in Fig. 5. Second, the model selection is greatly simplified. Rather than tuning the 4 model parameters simultaneously; the proposed model selection involves two steps, i.e.,

a. perform model selection for the C and kernel parameters for the standard SVM classifier (in step. 1a).
b. perform model selection for C*/C (ratio) while keeping C and kernel parameters fixed (as in (a)). Parameter ε is usually pre-set to a small value and does not require tuning. (For this paper we set $\varepsilon = 0$).

Cherkassky et al [3] demonstrate the utility of these conditions for several real-life data sets. Further, they establish connections between their practical conditions and the analytic results in [5]. However, like most other studies of the U-SVM, their paper assumes balanced data sets with equal misclassification costs. So there is a need to extend Universum learning to handle such cost-sensitive settings.

III. Cost-Sensitive Universum-SVM

Next we show how to introduce the misclassification costs directly into the U-SVM formulation (1). Consider a binary classification problem where the misclassification costs are given by the ratio $r = C_{fp} / C_{fn}$. This ratio r can be directly introduced into the U-SVM formulation (1) to handle cost-sensitive settings. This leads to the proposed modified cost-sensitive U-SVM formulation (2),

$$\min_{\mathbf{w},b} R(\mathbf{w},b) = \frac{1}{2}(\mathbf{w}\cdot\mathbf{w}) + C\sum_{i\in +class}\xi_i + \mathrm{C}\sum_{i\in -class} r\xi_i + C^*\sum_{j=1}^{m}\xi_j^* \qquad (2)$$

subject to constraints:

$$(\textit{training samples}):\ y_i[(\mathbf{w}\cdot\mathbf{x}_i)+b] \ge 1-\xi_i$$

$$(\textit{universum samples}):\ \left|(\mathbf{w}\cdot\mathbf{x}_j^*)+b\right| \le \varepsilon + \xi_j^*$$

$$\xi_i \ ,\ge 0, i=1,\dots,n \qquad \xi_j^* \ge 0, j=1,\dots,m$$

Here parameters r and $\varepsilon \ge 0$ are user-define. For this paper we set $\varepsilon = 0$ .Tunable regularization parameters $C, C^* \ge 0$ control the trade-off between minimization of cost-weighted errors, margin size and the maximization of the number of contradictions.

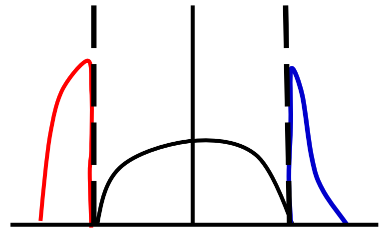

Fig.5. A schematic illustration of the histogram of projections of training and universum samples onto normal **w** vector of SVM decision boundary satisfying the practical conditions for the effectiveness of U-SVM.

Table 2. Practical Conditions for Effectiveness of U-SVM [3]

A1.	The histogram of projections of training samples is separable, and its projections cluster outside the SVM margin borders denoted as points -1/+1 in the projection space.
The histogram of projections of the Universum data:	
A2.	is symmetric relative to the (standard) SVM decision boundary, and
A3.	It has wide distribution between the SVM margin borders.

The proposed cost-sensitive U-SVM uses unequal costs for the two classes in the labeled training data, following [14, 15]. The samples of the negative class lying inside the soft-margin are penalized r times more than those of the positive class. However, the loss for the Universum samples remains the same as in the original formulation (1).

The optimization problem (2) can be solved by modifying the original U-SVM software [16], where the penalty term C for the negative class is multiplied by r . Hence, the computational cost for solving the cost-sensitive U-SVM problem remains the same as the standard U-SVM; which in turn is equivalent to solving the standard SVM problem with n+2m samples [2]. The modified cost-sensitive U-SVM software is made publicly available [17]. The solution to the optimization problem (2) defines the large margin hyper-plane $f(\mathbf{x}) = (\mathbf{w}^* \cdot \mathbf{x}) + b^*$ that incorporates a priori knowledge (i.e., Universum samples) and different misclassification costs into the final model.

As evident from (2), the cost-sensitive U-SVM suffers from the same problems as the original U-SVM, i.e., *model selection* and *selection of good Universum*. Hence, we adopt the same strategy as discussed in Table 1 for standard U-SVM. However, now the univariate histogram is generated by projecting the training and universum samples onto the normal direction vector of the *cost-sensitive SVM* hyperplane. Based on this histogram of projections, the practical conditions for the effectiveness of cost-sensitive U-SVM are provided in Table 3 and illustrated in Fig. 6. These new conditions (B1)-(B3) take into account the inherent 'bias' in the estimated predictive models under cost-sensitive settings, discussed in [18]. For equal cost settings, these conditions are equivalent to (A1)-(A3). Next, we provide empirical results to illustrate the proposed conditions for the effectiveness of cost-sensitive U-SVM.

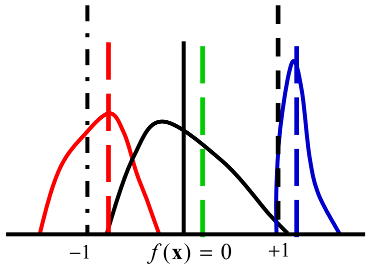

Fig.6. A schematic illustration of the histogram of projections onto normal **w** vector of cost-sensitive SVM decision boundary satisfying the practical conditions for the effectiveness of cost-sensitive U-SVM (when $r < 1$). Dashed **red**/**blue** lines indicate the training samples' class means. The average value of the two class means is shown in dashed **green**.

TABLE 3. PRACTICAL CONDITIONS FOR EFFECTIVENESS OF COST-SENSITIVE U-SVM

B1. The histogram of projections of the training data is well separable, and the samples from the class with smaller misclassification cost, (i.e. '+'ve class when $r < 1$) cluster outside the '+1' soft-margin.

Conditions for the histogram of projections of the Universum data:

B2. is *slightly* biased towards the class for which the misclassification cost is higher, (i.e. '–' ve class when $r < 1$), and

B3. is well spread *within* the class means of the training samples.

IV. EMPIRICAL RESULTS FOR COST-SENSITIVE U-SVM

This section presents two examples of applying the conditions (B1)-(B3) to determine the effectiveness of a given Universum dataset under cost-sensitive settings.

The *first set of experiment* uses *the synthetic 1000-dimensional hypercube data set*, where each input is uniformly distributed in [0, 1] interval and only 200 out of 1000 dimensions are relevant for classification. An output class label is generated as $y = \text{sign}(x_1+x_2+\ldots+x_{200} - 100)$. For this data set, only linear SVM is used because the optimal decision boundary is known to be linear. The training set size is 1,000, validation set size is 1,000, and test set size is 1,000. For U-SVM, 1,000 Universum samples are generated synthetically by using the commonly used strategy called *Random Averaging* (RA) [2, 3, 4]. For RA, the Universum samples are generated by *randomly* selecting positive and negative training samples, and computing their average.

For this data we use three different cost ratios r=0.5, 0.2, 0.1 to capture the effect of varying cost settings. We model this data for the standard/ cost-sensitive SVM and cost sensitive U-SVM using linear kernel. The model selection is performed by selecting the parameters with the smallest normalized weighted error $= \dfrac{r \times n_{fp} + n_{fn}}{r \times n^- + n^+}$ on

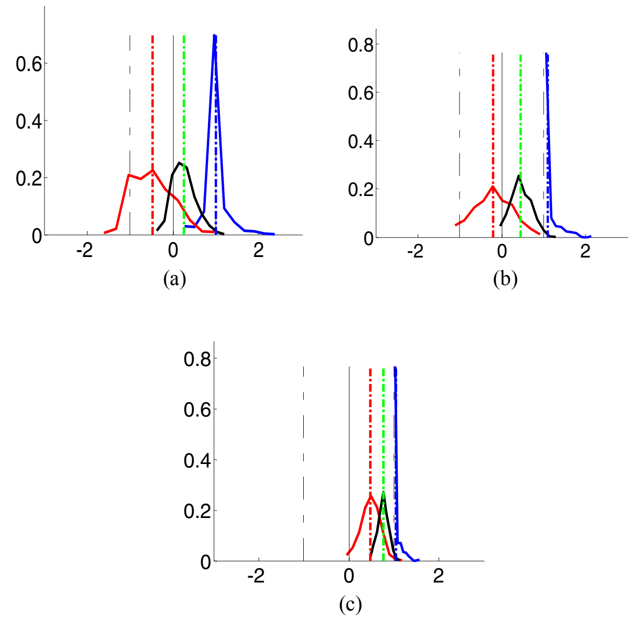

Fig. 7. Univariate histogram of projections onto cost-sensitive SVM normal weight vector for different cost-ratios (a) with k=0.5 ($C=2^{-6}$ and $C^*/C=2^{-4}$) (b) with k=0.2 ($C=2^{-5}$ and $C^*/C=2^{-8}$) (c) with k=0.1 ($C=2^{-5}$ and $C^*/C=2^{-5}$).

TABLE 4. COMPARISON OF STANDARD/COST-SENSITIVE SVM AND COST-SENSITIVE U-SVM FOR SYNTHETIC DATA

METHODS	standard SVM	cost-sensitive SVM	cost-sensitive U-SVM (RA)
Cost-Ratio r=0.5			
test error (in %)	**27.81(1.86)**	**24.84(1.38)**	**25.15(1.14)**
FP rate (in %)	27.49(8.57)	42.26(6.03)	39.9(5.72)
FN rate (in %)	27.96(6.39)	16.07(4.27)	17.69(3.74)
Cost-Ratio r=0.2			
test error (in %)	**21.21(5.68)**	**15.09(0.67)**	**14.92(0.57)**
FP rate (in %)	61.01(37.66)	73.75(14.07)	72.23(12.03)
FN rate (in %)	13.34(14.09)	3.37(2.26)	3.47(2.18)
Cost-Ratio r=0.1			
test error (in %)	**15.48(8.68)**	**8.80(0.43)**	**8.93(0.74)**
FP rate (in %)	68.79(37.22)	96.25(9.83)	90.99(11.53)
FN rate (in %)	10.24(13.17)	0.27(0.8)	0.93(1.52)

the independent validation set. Here, n_{fp}, n_{fn} denotes the number of false positive and false negative samples; and n^+, n^- denotes the number of positive and negative samples. Note that, the normalized weighted error is the weighted error ($C_{fp}n_{fp} + C_{fn}n_{fn}$) normalized by its maximum possible value i.e., $C_{fp}n^- + C_{fn}n^+$ [11,13,14].

Table 4 shows performance comparison for the standard SVM, cost-sensitive SVM and the cost-sensitive U-SVM with different cost-ratios (r=0.5, 0.2, 0.1). The table shows the average value of the normalized weighted test error over 10 random experiments. Here, for each experiment we randomly select the training / validation set, but use the same test set. The standard deviation of the weighted test error is provided in parenthesis. Additionally we provide the average False Positive and False Negative rates (in %) over these 10 random experiments. The typical histograms of projections for training data along with the Universum data are shown in Fig. 7. In all figures the training samples for the two classes are shown in red and blue with their respective class means shown by the dotted red/blue line. The projection of the universum samples are shown in black. Further, we also show the average of the two class means of the training samples in green. This helps to understand a projection bias of the universum samples towards positive or negative class. The typical histograms of projections in Fig.7 show that the training samples are not separable. Hence, based on our conditions we expect no improvement over the cost-sensitive SVM. This is confirmed by results in Table 4. For this data set (with unequal costs) application of standard Universum-SVM yields poor generalization (relative to cost-sensitive SVM).

The *next experiment* uses the *Real-life ISOLET data set* [19], where the data samples represent speech signals of 150 subjects for the letters 'B' vs. 'V'. Here, each sample is represented by 617 features that include spectral coefficients, contour features, sonorant features, pre-sonorant features, and post-sonorant features [19]. We label the voice signals for 'B' as class '+1' and 'V' as class '−1'. The cost-ratio is specified as,

$$r = \frac{C_{fp}}{C_{fn}} = \frac{\text{missclassification cost(truth='V',prediction='B')}}{\text{missclassification cost(truth='B',prediction='V')}}$$

For this experiment we use,

– No. of Training samples= 100. (50 per class)
– No. of Universum samples = 300 (2 types of Universa: letters P and RA).
– No. of Test samples=500. (This independent test set is used for model selection)

Initial experiments proved linear SVM to work well for this ISOLET dataset. Comparisons of the (linear) standard SVM, cost-sensitive SVM and the cost-sensitive U-SVM for the different types of Universa: letters P and RA with different cost-ratios (r=0.5, 0.2, 0.1) are shown in Table 5. The typical histograms of projections for training data along with the Universum data are also shown in Figs 8, 9 and 10. From these figures, it is clear that the training samples are well-separable. Analysis of projections for different types of universum samples shows that:

– *letter* 'P' has *well spread* projections between the training samples' class means and *slightly biased* towards the '–' ve class.
– *Random Averaging* has *narrower* projections than the letter "P", and are *slightly biased* towards the '+' ve class.

Hence, we can expect letter "P" to be more effective than RA. This is consistent with the empirical results in Table 5.

TABLE 5. COMPARISON OF STANDARD/COST-SENSITIVE SVM AND COST-SENSITIVE U-SVM FOR REAL LIFE ISOLET DATA.

METHODS	standard SVM	cost-sensitive SVM	cost-sensitive U-SVM (letter P)	cost-sensitive U-SVM (RA)
Cost-Ratio (r=0.5)				
test error (%)	**5.34(1.47)**	**5.21(1.23)**	**4.33(0.82)**	**4.96(1.05)**
FP rate (%)	9.36(2.31)	10.20(4.08)	10.20(3.78)	9.52(3.40)
FN rate (%)	3.32(1.61)	2.72(1.90)	1.40(0.84)	2.68(1.85)
Cost-Ratio r=0.2				
test error (%)	**3.51(0.51)**	**3.42(0.42)**	**2.77(0.52)**	**3.03(0.53)**
FP rate (%)	11.68(3.20)	12.56(3.18)	13.6(3.76)	11.96(2.59)
FN rate (%)	1.88(0.98)	1.60(0.75)	0.60(0.43)	1.24(0.74)
Cost-Ratio (r=0.1)				
test error (%)	**2.79(0.75)**	**2.70(0.65)**	**1.78(0.42)**	**2.39(0.69)**
FP rate (%)	12.24(3.81)	15.28(4.28)	17.6(4.82)	14.6(3.93)
FN rate (%)	1.84(0.76)	1.44(0.66)	0.2(0.28)	0.48(0.45)

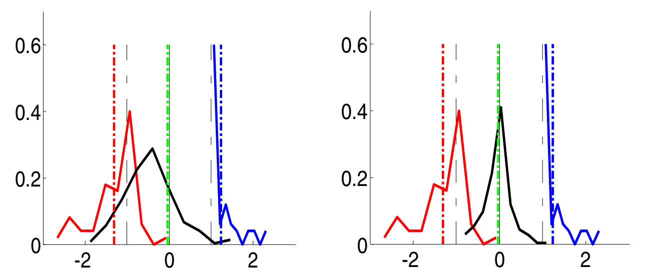

Fig. 8. Univariate histogram of projections onto cost-sensitive SVM normal weight vector ($C=2^{-4}$) for different types of Universa for k=0.5. (a) letter P Universum. $C^*/C=2^{-5}$ (b) RA Universum $C^*/C=2^{-4}$.

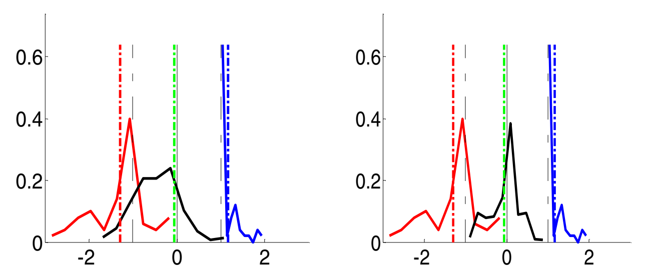

Fig. 9. Univariate histogram of projections onto cost-sensitive SVM normal weight vector ($C=2^{-4}$) for different types of Universa for k=0.2. (a) letter P Universum. $C^*/C=2^{-5}$ (b) RA Universum $C^*/C=2^{-2}$.

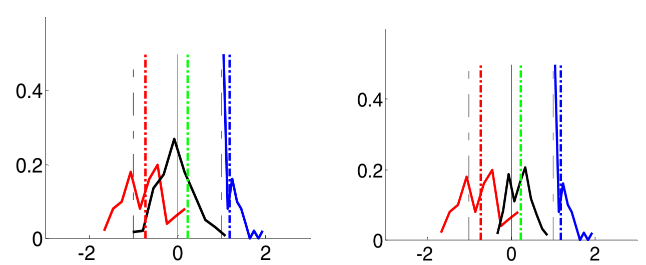

Fig. 10. Univariate histogram of projections onto cost-sensitive SVM normal weight vector ($C=2^{-3}$) for different types of Universa for k=0.1. (a) letter P Universum. $C^*/C=2^{-6}$ (b) RA Universum $C^*/C=2^{-5}$.

Additional empirical results using several other real-life datasets including *MNIST handwritten digit recognition dataset* [20], and *German Traffic Sign Recognition Benchmark dataset* [21] further confirm the efficiency of the practical conditions in Table 3. The results have not been included due to space constraints.

V. CONCLUSIONS

The idea of using Universum to improve SVM based learning in high-dimensional finite sample settings have been widely studied in [2-10]. However, most of these studies use balanced data sets with equal misclassification costs. This paper proposes a new formulation to incorporate cost-sensitive learning into the U-SVM optimization. The cost-sensitive U-SVM formulation can be implemented using minor modifications to standard U-SVM software. This modified software is made publicly available at [17].

We presented practical conditions for the effectiveness of the cost-sensitive U-SVM using histograms of projections. These conditions can be easily applied by general users, because:

1. They provide an explicit characterization of the properties of the Universum relative to the properties of labeled training data. These properties are conveniently represented in the form of the univariate histograms of projections;
2. They directly relate prediction performance of cost-sensitive U-SVM to that of cost-sensitive SVM.

Further, the proposed approach significantly simplifies model selection for the cost-sensitive U-SVM. That is, the regularization parameter C and the kernel parameter for the cost-sensitive U-SVM formulation (2) are selected via training a cost-sensitive SVM classifier. Then model selection for cost-sensitive U-SVM involves tuning only one parameter, C*/C.

Finally, we point out that many applications involve extreme scenarios with very high cost ratios or extreme unbalance in the data (viz., anomaly detection). The typical histograms under such extreme scenarios may lose the property of the separability of the training samples. Hence, there is a need for future research on the effectiveness of Universum under such extreme settings.

ACKNOWLEDGMENT

This work was supported, in part, by NSF grant ECCS-0802056. The authors gratefully acknowledge several discussions with Prof. Vladimir Vapnik from Columbia University on the Universum SVM learning.

REFERENCES

[1] V. N. Vapnik*, Estimation of Dependencies Based on Empirical Data: Empirical Inference Science:* Afterword of 2006. New York: Springer- Verlag, 2006.

[2] J. Weston,R. Collobert,F. Sinz,L. Bottou, and V. Vapnik, “Inference with the Universum,” *Proc. ICML*,2006, pp. 1009–1016.

[3] V. Cherkassky, S. Dhar, and W. Dai, “Practical Conditions for Effectiveness of the Universum Learning,”, *IEEE Transactions on Neural Networks*,vol.22, no. 8, pp. 1241-1255, Aug 2011.

[4] V. Cherkassky, and W. Dai, “Empirical Study of the Universum SVM Learning for High-Dimensional Data,” *in Proc. ICANN*, 2009.

[5] F. Sinz, O. Chapelle, A. Agarwal, and B. Schölkopf, “An analysis of inference with the Universum,” *In Proc. of 21st Annual Conference on Neural Information Processing Systems*, 2008, pp. 1–8.

[6] T. T. Gao, Z.X Yang, L. Jing, "On Universum-Support Vector Machines",The Eighth International Symposium on Operations Research and Its Applications, China, 2009, pp. 473-480.

[7] D. Zhang, J. Wang, F. Wang, and C. Zhang, “Semi-supervised classification with Universum,” *Proceedings of the 8th SIAM Conference on Data Mining (SDM)*, 2008, pp. 323–333.

[8] S. Chen and C. Zhang, “Selecting informative Universum sample for semi-supervised learning,” *in Proc. Int. Joint Conf. Artif. Intell.*, 2009, pp. 1016–1021.

[9] X. Bai and V. Cherkassky, “Gender classification of human faces using inference through contradictions,” *in Proc. Int. Joint Conf. Neural Netw.*, Hong Kong, Jun. 2008, pp. 746–750.

[10] C. Shen, P. Wang, F. Shen, H. Wang, "UBoost: Boosting with the Universum", IEEE Transaction on Pattern Analysis and Machine Intelligence, 2011.

[11] P.N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining.* New York: Pearson Education, 2006.

[12] G. M. Weiss, K. McCarthy, B. Zabar,“Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?”, *DMIN* 2007,pp. 35-41.

[13] C. Elkan, “The foundations of cost-sensitive learning”, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001.

[14] V. Cherkassky and F. Mulier, *Learning from Data Concepts: Theory and Methods*, 2nd ed. New York: Wiley, 2007.

[15] Y. Lin, Y. Lee, and G. Wahba, "Support vector machines for classification in nonstandard situations", *Machine Learning*, vol. 46,pp. 191-202, 2002.

[16] UniverSVM. [WWW page].URL: http://mloss.org/software/view/19/.

[17] Cost Sensitive Univerum Software. [WWW page]. URL: http://www.ece.umn.edu/users/cherkass/predictive_learning/SOFTWARES.html.

[18] V. Cherkassky and S. Dhar, “Simple method for interpretation of high dimensional nonlinear SVM classification models,” *in Proc. Int. Conf. Data Min.*, Las Vegas, NV, Jul. 2010, pp. 267–272.

[19] M. Fanty, R. Cole. “Spoken letter recognition”, *Advances in Neural Information Processing Systems* 3. San Mateo, CA: Morgan Kaufmann, 1991.

[20] S. Roweis, sam roweis: data. [WWW page]. URL http://www.cs.nyu.edu/~roweis/ data.html.

[21] The German Traffic Sign Recognition Benchmark. [WWW page]. URL http://benchmark.ini.rub.de/?section=gtsrb&subsection=dataset#resultanalysis.