Development and Evaluation of Cost-Sensitive Universum SVM

Sauptik Dhar and Vladimir Cherkassky, *Fellow, IEEE.*

Abstract— Many machine learning applications involve analysis of high-dimensional data, where the number of input features is larger than/comparable to 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 Contradiction or Universum support vector machine (U-SVM) [1, 2]. Recent studies [2-10] have shown U-SVM to be quite effective for sparse high-dimensional data sets. However, all these earlier studies have used balanced data sets with equal misclassification costs. This paper extends the U-SVM formulation to problems with different misclassification costs, and presents practical conditions for the effectiveness of this cost-sensitive U-SVM. Several empirical comparisons are presented to validate the proposed approach.

Index Terms— Cost-sensitive SVM, learning through contradiction, misclassification costs, Universum SVM.

1 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 the Universum learning to be particularly effective for high-dimensional 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 importance (or "cost") of false positive and false negative errors is assumed to be the same. However, many practical applications involve unbalanced data and unequal misclassification costs. Examples include credit card fraud detection, intrusion detection, oil-spill detection, medical diagnosis etc. [11-13]. In order to incorporate a priori knowledge (in the form of Universum data), we need to extend the Universum learning to handle such cost-sensitive settings.

Researchers have introduced many techniques to deal with unequal misclassification costs and unbalanced data settings [11-13]. Typically, these methods follow two basic approaches:

– *Cost-Sensitive Learning*, where the costs of misclassification and the ratio of imbalance in the data are introduced directly into the learning formulation [14-16].

– *Sampling-based approaches*, where the training samples of a particular class are replicated to reflect unequal misclassification costs [13]. Such strategies exploit the equivalency between changing the proportion of positive and negative training samples and the misclassification costs [13]. There are three sampling approaches:

a. *Oversampling* replicates samples (of the minority class) until training data has equal number of positive and negative samples or equal misclassification costs.

b. *Undersampling* removes samples (of the majority class) until training data has equal number of positive and negative samples or equal misclassification costs.

c. *Hybrid* methods use a combination of undersampling and oversampling to achieve more balanced class distribution and/or equal misclassification costs.

Note that cost-sensitive learning enables better analytic understanding, while sampling-based methods are usually adopted by practitioners. This paper follows the direct approach of introducing the cost-ratios into Universum-SVM formulation. Specifically, we introduce the U-SVM classification setting, where different misclassification costs for *false-positive* vs. *false-negative* errors are given as the ratio $r = C_{fp}/C_{fn}$. We extend our work presented in [14] and modify 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. Our approach follows a practical strategy that aims to answer two practical questions:

i. Can a particular Universum data set improve generalization performance of the cost-sensitive SVM classifier [15, 16] trained using only labeled data?

• *Sauptik Dhar is with the Research and Technology Center, Robert Bosch LLC, Palo Alto, CA 94304. Email: sauptik.dhar@us.bosch.com*

• *Vladimir Cherkassky is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis MN 55455. E-mail: cherk001@umn.edu*.

Manuscript received (insert date of submission if desired).

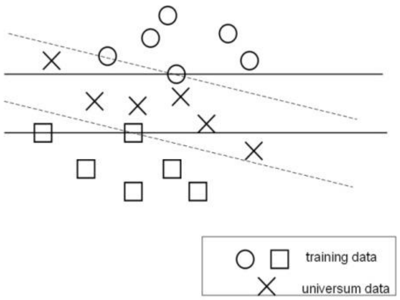


Fig.1. Two large-margin separating hyperplanes explain training data equally well, but have different number of contradictions on the Universum. The model with a larger number of contradictions should be selected.

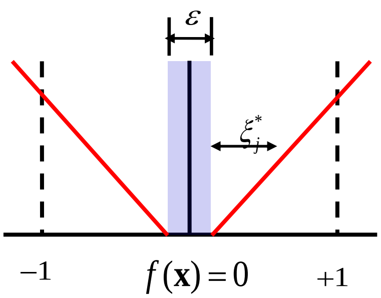


Fig. 2. The ε - insensitive loss for the Universum samples. Universum samples outside the ε -insensitive zone are linearly penalized using the slack variables ξ_j^* .

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 standard cost-sensitive SVM. Our conditions for the effectiveness of cost-sensitive U-SVM extend conditions for the effectiveness of the standard U-SVM introduced in [3].

The paper is organized as follows. Section 2 describes Vapnik's original formulation for U-SVM [1] and presents practical conditions for its effectiveness [3]. Section 3 presents new cost-sensitive U-SVM formulation and the practical conditions for its effectiveness. Section 4 provides empirical results to illustrate these conditions, using both synthetic and real-life data sets. Finally, conclusions are presented in Section 5.

2 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. The Universum learning was introduced 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 unlabeled Universum samples are incorporated into learning as explained next. Let us assume that labeled training data is linearly separable using large-margin hyperplane. Then the Universum samples can fall either inside or outside the margin borders (see Fig. 1). Under U-SVM, we favor large-margin 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).

Next, we briefly review the optimization formulation for Universum SVM classifier [1, 2]. Let us consider an inductive setting (for binary classification), where we have labeled training data $(\mathbf{x}_i, y_i),\ i = 1, 2, ...n$ and a set of unlabeled examples $(\mathbf{x}_j^*),\ j = 1, 2, ...m$ from the Universum. The analytic formulation for U- SVM [1, 2] is shown in Box (1). Note that all SVM optimization formulations in this paper are presented only for linear parameterization; but they can be

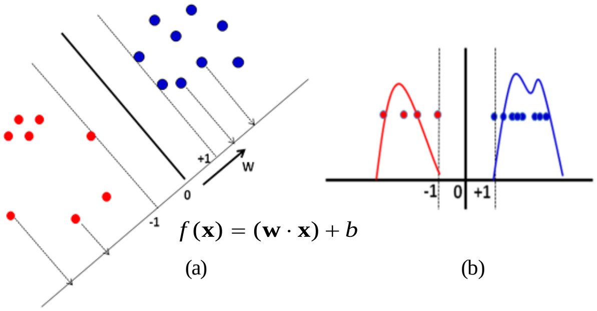


Fig. 3. (a) Projection of the training data shown in red and blue onto the normal weight vector ($\mathbf{w}$) of the SVM hyperplane. (b) Univariate histogram of projections. i.e. histogram of $f(\mathbf{x})$ values for training samples.

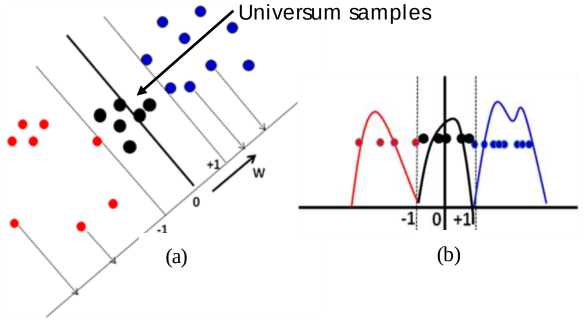


Fig. 4. Histogram of projections technique.
(a) Projection of the universum data (shown in **black)** onto the normal weight vector ($\mathbf{w}$) of the SVM hyperplane.
(b) Histogram of projections of the universum samples (shown in **black**) along with the training samples (shown in red/blue).

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).

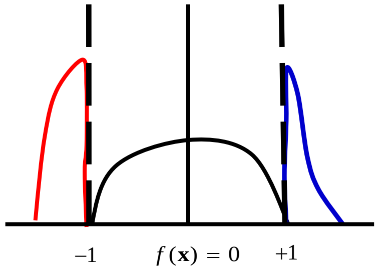


Fig.5. A schematic illustration of the histogram of projections of training and universum samples onto normal **w** vector of SVM decision boundary satisfying the practical conditions for the effectiveness of U-SVM.

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 SVM margin borders.

readily extended to nonlinear case using kernels. Here, for labeled training data, we use standard SVM soft-margin loss with slack variables ξ_i. The Universum samples $(\mathbf{x}^*_j)$ are penalized via ε–insensitive loss (shown in Fig. 2). Let ξ^*_j denote slack variables for 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:

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

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

$$\xi_i \geq 0,\; i=1,\ldots,n \qquad \xi^*_j \geq 0,\; j=1,\ldots,m$$

Here parameter $\varepsilon \geq 0$ is user-defined and usually set to zero or a small value. Parameters $C, C^* \geq 0$ control the trade-off between the margin size, the number of errors and the number of contradictions. Note that for $C^* = 0$ this formulation becomes equivalent to standard SVM classifier [15].

The solution to the optimization problem (1) yields a large-margin hyperplane that also incorporates a priori knowledge (i.e., Universum data) into the final model. There are *two design factors* important for successful application of U-SVM:

– *Model Selection*: which becomes rather difficult because the kernelized U-SVM has 4 tuning parameters: C, C^*, kernel parameter and ε (vs. two parameters in standard SVM).

– generalization performance of U-SVM may be negatively affected by a poor choice of the Universum data.

In practice, it may be difficult to separate these two factors. The strategy for judging the effectiveness of a given Universum is described in [3]. This strategy 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 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. Practical conditions for the effectiveness of U-SVM [3] are provided in Table 2 and illustrated in Fig. 5. The second aspect of the proposed strategy is simplified model selection. Specifically, this strategy involves two steps, i.e.,

a. First, perform optimal tuning of the C and kernel parameters for standard SVM classifier (in step 1a).

b. Second, perform tuning of the ratio C*/C, 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.

Cherkassky et al [3] demonstrate the effectiveness 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 all 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.

3 COST-SENSITIVE UNIVERSUM-SVM

Consider a binary classification problem where we have labeled training samples and unlabeled Universum samples, as in standard U-SVM described in Section 2. However, we assign different importance (or cost) to false positive and false negative errors, as specified by the ratio $r = C_{fp}/C_{fn}$. The goal of cost-sensitive learning is to estimate a classifier that minimizes the weighted error $C_{fp}P_{fp} + C_{fn}P_{fn}$ for future test samples [13, 15, 16]. Here P_{fp} and P_{fn} denote the probability (error rate) of false positive and false negative errors. For empirical comparisons, this weighted test error is normalized by its maximum possible value ($C_{fp} + C_{fn}$), as shown next:

$$\textit{Normalized}\,(C_{fp}P_{fp} + C_{fn}P_{fn}) = \frac{r\times(n_{fp}/n^-)+(n_{fn}/n^+)}{r\,(n^-/n^-)+(n^+/n^+)}$$

$$= \frac{r\times(n_{fp}/n^-)+(n_{fn}/n^+)}{r+1}$$

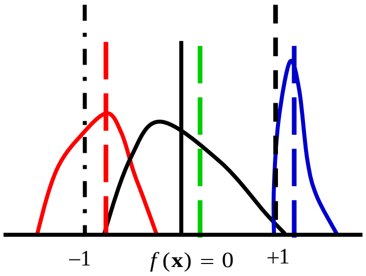


Fig.6. A schematic illustration of the histogram of projections onto normal **w** vector of cost-sensitive SVM decision boundary satisfying the practical conditions for the effectiveness of cost-sensitive U-SVM (when $r < 1$). Dashed red/blue lines indicate the training samples' class means. The average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

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. Such normalization limits the value of the weighted error to the range of [0, 1], which is the same range used in standard binary classification problems (with equal costs). In the the rest of the paper, we refer to this *normalized weighted test error* as simply the *test error*.

Several alternative metrics have been used in literature to measure the performance of a classification model under unbalanced and unequal misclassification costs settings [11, 15]. This paper advocates using cost-sensitive U-SVM only if it provides an improvement over standard cost-sensitive SVM [15, 16]. Following [16], it has been shown that the minimizer for the expected value of the loss function for the cost-sensitive SVM follows the Bayes rule. This provides theoretical justification for using an empirical estimate of the Bayes Risk (i.e., the weighted test error) for empirical comparisons presented in Section 4.

Next, we present an extension of the Universum learning to cost-sensitive settings. As discussed in Section 1, there exist several approaches for handling cost-sensitive settings [11-13]. This paper follows the direct approach of introducing the cost-ratio $r = C_{fp}/C_{fn}$ directly into the U-SVM formulation (1). This leads to the modified cost-sensitive U-SVM formulation shown in Box 2.

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

subject to constraints:

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

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

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

Here, parameters r and $\varepsilon \ge 0$ are user-defined. In all empirical results presented in Section 4, the value of ε is set to zero. 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.

The proposed cost-sensitive U-SVM uses unequal costs for the two classes in the labeled training data, following [15, 16]. 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). Note that when $C^* = 0$ this formulation is equivalent to standard cost-sensitive SVM [15, 16].

Following [2], this quadratic optimization problem (2) can be solved by introducing the Univerum samples twice with opposite labels and hence solving a modified cost-sensitive SVM problem.
That is, we introduce

$$\mathbf{x}_{n+j} = \mathbf{x}_j^* \text{ and } y_{n+j} = +1, \; j = 1,2,...m$$

$$\mathbf{x}_{n+j} = \mathbf{x}_j^* \text{ and } y_{n+j} = -1, \; j = m+1, m+2,...2m$$

Then (2) is equivalent to solving the following optimization problem,

$$\min_{\mathbf{w},b} R(\mathbf{w},b) = \frac{1}{2}(\mathbf{w}\cdot\mathbf{w}) + \hat{C}\sum_{i=1}^{n+2m} k_i\xi_i \qquad (3)$$

subject to constraints: $y_i[(\mathbf{w}\cdot\mathbf{x}_i)+b] \ge \rho_i - \xi_i$

$\xi_i \ge 0, \qquad i = 1,...,n+2m$

where,

$\rho_i = 1$ and $\hat{C} = C$; for $i = 1,...,n$

$\rho_i = -\varepsilon$ and $\hat{C} = C^*$; for $i = n+1,...,n+2m$

and, $k_i = \begin{cases} C_{fp}/C_{fn} & \text{if } y_i = -1 \; (i = 1,2,...\,n) \\ 1 & \text{otherwise} \end{cases}$

This problem (3) can be easily solved in the dual form by using the original U-SVM software [17] where the $\hat{C}$ penalty term for the negative samples is weighted by the factor

$r = C_{fp} / C_{fn}$. Hence, the computational cost for solving the cost-sensitive U-SVM problem remains the same as for the standard U-SVM, which is in turn equivalent to solving the standard SVM problem with *n*+2*m* samples [2]. The modified cost-sensitive U-SVM software is made publicly available [18]. The solution to the optimization problem (2) defines the large margin hyper-plane $f(x) = (\mathbf{w}^* \cdot \mathbf{x}) + b^*$ that incorporates a priori knowledge (i.e., Universum samples) and also reflects different misclassification costs.

As evident from the optimization formulation (2), cost-sensitive U-SVM has the same design issues as the original U-SVM, i.e., *model selection* and *selection of good Universum*. These issues can be addressed via the same general strategy as our earlier approach used for standard U-SVM (see Table 1). 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, new 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 SVM models under cost-sensitive settings [19, 20]. Conditions (A1)-(A3) represent a special case of conditions (B1)-(B3) when the costs are equal ($r = 1$). Further, we propose the following two-step strategy for model selection for the cost-sensitive U-SVM:

1. perform model selection for C and kernel parameters for the cost-sensitive SVM formulation. (These parameters are then fixed and used for the cost-sensitive U-SVM).
2. perform model selection for the C*/C parameter specific to the cost-sensitive U-SVM formulation, while keeping C and kernel parameters fixed. Parameter ε is usually pre-set to a small value and does not require tuning.

This strategy is used in all empirical comparisons reported in Section 4 below (where parameter ε is set to zero).

4 Empirical Results for Cost-Sensitive U-SVM

This section presents empirical results to illustrate the conditions (B1)-(B3) for the effectiveness of cost-sensitive Universum SVM.

The *first set of experiments* 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 = sign(x_1+x_2+…+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 from the training data using the *Random Averaging* (RA) strategy [2, 3, 4]. That is, Universum samples are generated by *randomly* selecting positive and negative training samples, and then computing their average.

For this data set, we consider 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 SVM, cost-sensitive SVM and cost sensitive U-SVM using linear kernel. The model selection is performed by tuning parameter values providing the smallest normalized weighted error on the independent validation set.

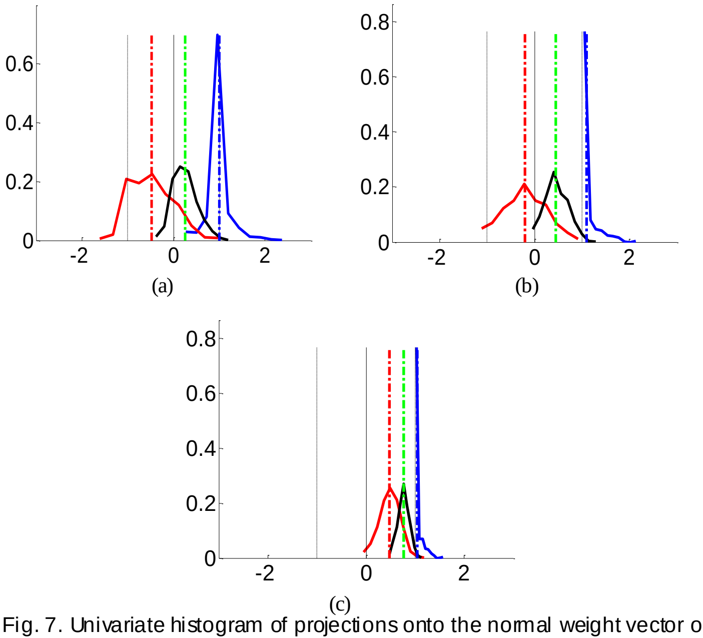


Fig. 7. Univariate histogram of projections onto the nomal weight vector of cost-sensitive SVM for different cost-ratios: (a) r=0.5(C=2^{-6} and C*/C=2^{-4}), (b) r=0.2 (C=2^{-5} and C*/C=2^{-8}), (c) r=0.1 (C=2^{-5} and C*/C=2^{-5}).

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)

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 test error is shown in parenthesis. Additionally we provide the average False Positive and False Negative test error rates over 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 indicated by the dotted

TABLE 5. COMPARISON OF STANDARD SVM, COST-SENSITIVE SVM AND COST-SENSITIVE U-SVM FOR REAL LIFE MNIST DATA (USING LINEAR KERNEL).

METHODS	standard SVM	cost-sensitive SVM	cost-sensitive U-SVM (digit 1)	cost-sensitive U-SVM(digit 3)	cost-sensitive U-SVM(digit 6)	cost-sensitive U-SVM(RA)
	Cost-Ratio (r=0.5)					
test error (%)	**4.80(0.51)**	**4.40(0.38)**	**4.39(0.31)**	**4.36(0.32)**	**4.33(0.44)**	**4.37(0.46)**
FP rate (in %)	3.94(0.50)	5.67(1.48)	5.64(1.35)	6.00(1.37)	5.84(1.40)	5.54(1.23)
FN rate (in %)	5.29(0.81)	3.82(0.69)	3.82(0.67)	3.60(0.67)	3.63(0.75)	3.84(0.67)
	Cost-Ratio (r=0.2)					
test error (%)	**4.91(0.48)**	**3.15(0.22)**	**3.12(0.24)**	**3.13(0.17)**	**3.17(0.21)**	**3.19(0.25)**
FP rate (in %)	3.92(0.58)	10.96(2.96)	11.10(2.90)	11.45(3.05)	11.38(2.71)	10.64(2.05)
FN rate (in %)	5.09(0.55)	1.72(0.47)	1.65(0.44)	1.60(0.50)	1.66(0.45)	1.83(0.56)
	Cost-Ratio (r=0.1)					
test error (%)	**5.03(0.72)**	**2.41(0.34)**	**2.36(0.33)**	**2.33(0.34)**	**2.31(0.30)**	**2.39(0.29)**
FP rate (in %)	4.57(0.72)	13.33(2.42)	13.94(2.88)	15.17(4.04)	14.54(3.48)	13.94(2.43)
FN rate (in %)	5.07(0.75)	1.41(0.51)	1.30(0.53)	1.15(0.50)	1.18(0.57)	1.33(0.47)

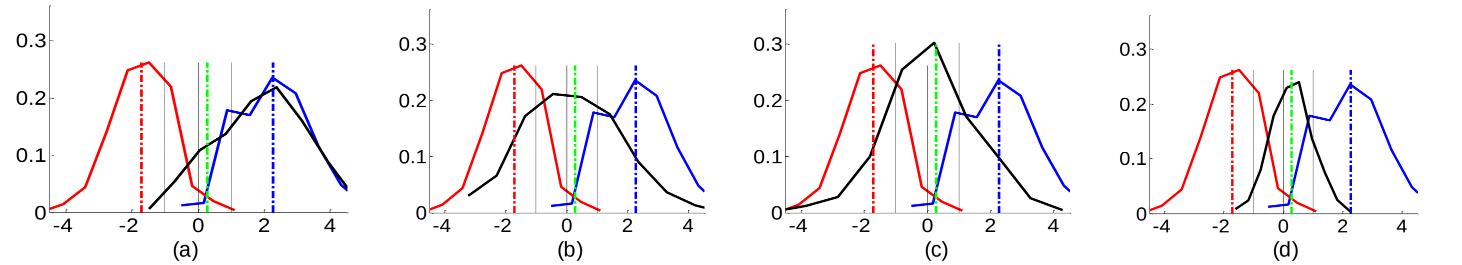


Fig. 8. Univariate histogram of projections onto the normal weight vector of cost-sensitive SVM (r=0.5, C=2^{-4}) for different types of Universa. Training set size ~1,000 samples. Universum set size ~1,000 samples. (a) Digit 1 Universum C*/C=2^{-9} (b) Digit 3 Universum C*/C=2^{-5} (c) Digit 6 Universum C*/C=2^{-8} (d) RA Universum C*/C=2^{-5}.

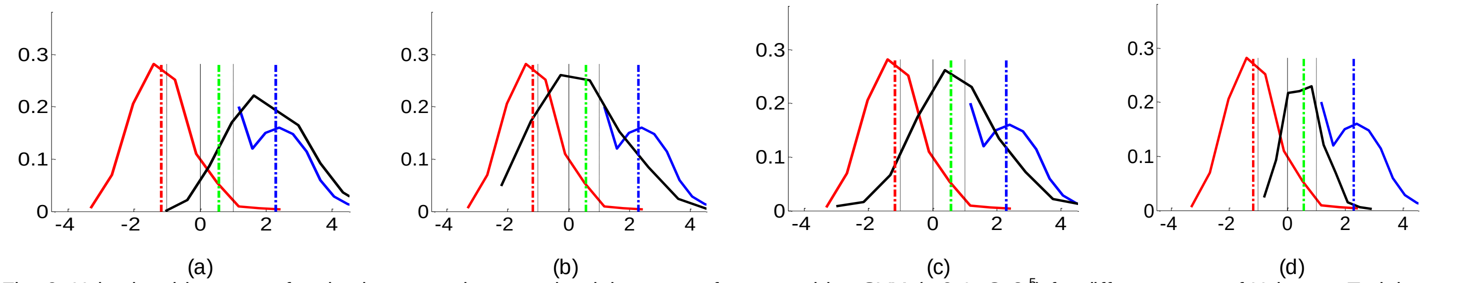


Fig. 9. Univariate histogram of projections onto the normal weight vector of cost-sensitive SVM (r=0.1, C=2^{-5}) for different types of Universa. Training set size ~1,000 samples. Universum set size ~1,000 samples. (a) Digit 1 Universum C*/C=2^{-20} (b) Digit 3 Universum C*/C=2^{-7} (c) Digit 6 Universum C*/C=2^{-10} (d) RA Universum C*/C=2^{-7}.

red/blue line. The histogram of projections of the universum samples is shown in black. Further, we also show the average of the two class means of the training samples in green. This helps to illustrate the projection bias of the universum samples towards positive or negative class. Typical histograms of projections (in Fig. 7) show that the training samples are not separable. Hence, according to condition B1 (in Table 3), we expect no improvement over the cost-sensitive SVM. This is consistent with results in Table 4. For this data set (with unequal costs), introducing Universum does not improve generalization (relative to standard cost-sensitive SVM).

The *second set of experiments* uses handwritten digits "8" vs. "5" MNIST data [21]. The goal is accurate classification of digits "8" vs. "5", where each sample is represented as a real-valued vector of size 28x28=784. We use four types of Universa: handwritten digits "1", "3", "6" and RA and analyze their effectiveness using the histograms of projections of both labeled and Universum data sets. For this experiment,

- Number of training samples ~ 1000 (500 per class).
- Number of validation samples ~ 1000 (500 per class. This independent validation set is used for model selection).
- Number of test samples ~1866 (i.e., 892 samples of digit "8" and 974 samples of digit "5").
- Number of Universum samples ~ 1000.
- Linear SVM parameterization is used.

Digit "8" samples correspond to a positive class and digit "5" to negative class. So misclassification costs are defined as:

$$\frac{\text{missclassification cost for(truth=digit 5,prediction=digit 8)}}{\text{missclassification cost for(truth=digit 8,prediction=digit 5)}} = \frac{C_{fp}}{C_{fn}} = r$$

Table 5 shows performance comparisons between standard SVM, cost-sensitive SVM and the cost-sensitive U-SVM for different types of Universa (digit “1”, “3”, “6” and RA) and for different cost-ratios (r=0.5, 0.2, 0.1). Typical histograms of projections for training data along with the Universum data are shown in Figs. 8 and 9. For this data set the histograms of projections for the cost-ratio r=0.2 are not shown, because they look very similar to those for r=0.1. Visual analysis of these histograms indicates that the training samples are not separable; hence, cost-sensitive U-SVM is not likely to provide any improvement over the cost-sensitive SVM. This is consistent with empirical results shown in Table 5.

Standard sampling-based approaches are technically equivalent to setting different misclassification costs [15, 16]. For example, consider the above experiment for the cost-sensitive SVM with r = 0.5. A typical oversampling solution approach would use the training data set containing 1,000 positive samples (of digit ‘8’) and 500 negative samples (of digit ‘5’) to estimate a standard SVM classifier (with equal misclassification costs). This is equivalent to using a penalty of 2C for the positive class and C for the negative class (in the formulation (1) with C*=0). Of course, this oversampling approach is mathematically equivalent to solving the cost-sensitive SVM formulation with r=0.5 (see formulation (2) with C*=0). For this current experiment with r = 0.5, the oversampling solution approach yields a test error of 4.47 % with FP rate of 5.88 % and FN rate of 3.82 %. These results are practically the same as error rates shown in Table 5 (obtained via cost-sensitive solution approach). Detailed theoretical analysis of the equivalence between cost-sensitive and the sampling-based approaches can be found in [13].

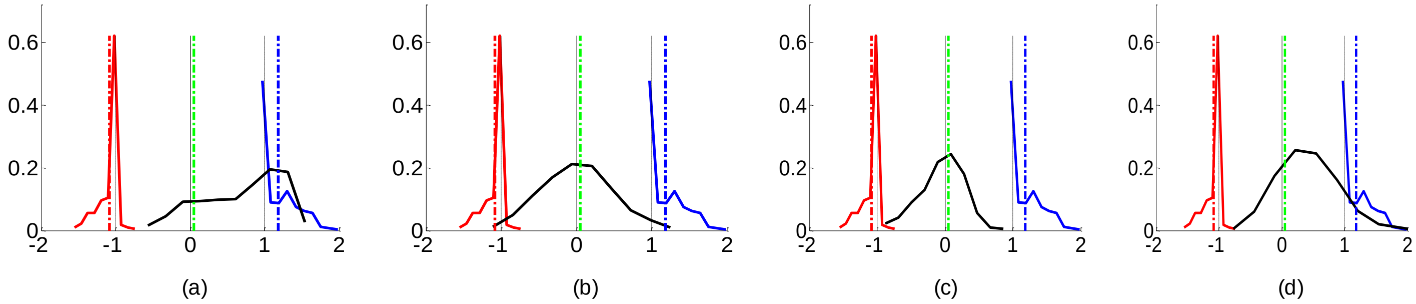


Fig. 10. Univariate histograms of projections for cost-sensitive SVM with r=0.5 (C=2, $\gamma = 2^{-6}$), for different types of Universa. Training set size ~1000 samples. Universum set size ~1000 samples. (a) Digit 1 Universum C*/C=2^{-4}. (b) Digit 3 Universum C*/C=2^{-2}. (c) Digit 6 Universum C*/C=2^{-2}. (d) RA Universum C*/C=2^{-1}.

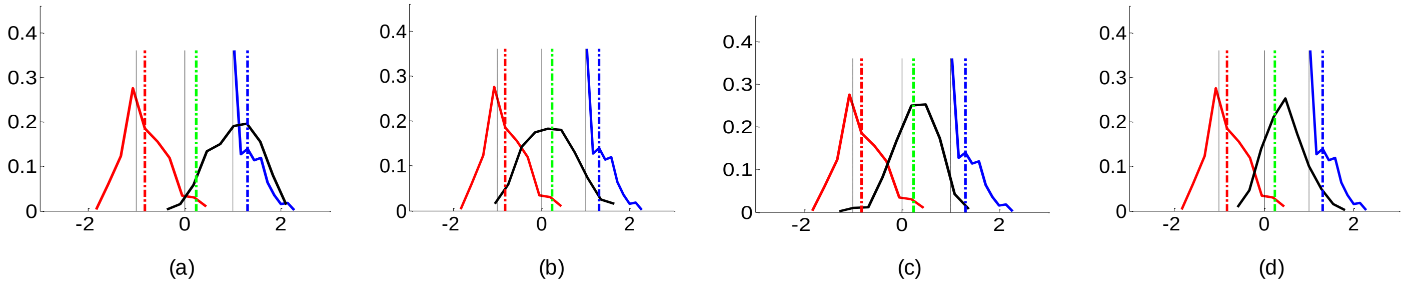


Fig. 11. Univariate histograms of projections for cost-sensitive SVM with r=0.1 (C=2, $\gamma = 2^{-6}$), for different types of Universa. Training set size ~1000 samples. Universum set size ~1000 samples. (a) Digit 1 Universum C*/C=2^{-4}. (b) Digit 3 Universum C*/C=2^{-4}. (c) Digit 6 Universum C*/C=2^{-4}. (d) RA Universum C*/C=2^{-2}.

TABLE 6. COMPARISON OF STANDARD SVM, COST-SENSITIVE SVM AND COST-SENSITIVE U-SVM FOR REAL LIFE MNIST DATA (USING RBF KERNEL).

METHODS	standard SVM	cost-sensitive SVM	cost-sensitive U-SVM (digit 1)	cost-sensitive U-SVM(digit 3)	cost-sensitive U-SVM(digit 6)	cost-sensitive U-SVM (RA)
	Cost-Ratio (r=0.5)					
test error (%)	**1.34(0.28)**	**1.31(0.29)**	**1.23(0.37)**	**0.95(0.19)**	**1.15(0.34)**	**1.16(0.28)**
FP rate (in %)	1.10(0.73)	1.12(0.72)	0.96(0.66)	1.07(0.82)	0.89(0.74)	1.03(1.12)
FN rate (in %)	1.45(0.29)	1.41(0.3)	1.35(0.36)	0.89(0.27)	1.27(0.35)	1.23(0.27)
	Cost-Ratio (r=0.2)					
test error (%)	**1.59 (0.25)**	**1.45(0.20)**	**1.29(0.28)**	**0.97(0.31)**	**1.11(0.22)**	**1.17(0.28)**
FP rate (in %)	1.15(0.24)	3.19 (2.26)	3.43(2.69)	3.35(2.71)	2.64(2.27)	3.00(3.48)
FN rate (in %)	1.67(0.32)	1.13(0.44)	0.90(0.50)	0.53(0.39)	0.83(0.47)	0.84(0.54)
	Cost-Ratio (r=0.1)					
test error (%)	**1.50(0.24)**	**1.13(0.19)**	**1.11(0.17)**	**0.80(0.14)**	**0.90(0.22)**	**0.92(0.17)**
FP rate (in %)	1.31(1.47)	5.91(2.75)	6.57(3.27)	6.29(3.20)	5.24(2.54)	6.58(3.62)
FN rate (in %)	1.52(0.28)	0.69(0.37)	0.61(0.33)	0.30(0.19)	0.51(0.27)	0.41(0.27)

The *3rd set of experiments* uses the same *real-life* handwritten digits "8" vs. "5". However, here we use an RBF kernel of the form $K(\mathbf{x},\mathbf{x}') = \exp\left(-\gamma\left\|\mathbf{x}-\mathbf{x}'\right\|^2\right)$. Table 6 shows empirical performance comparisons between standard SVM, cost-sensitive SVM and the cost-sensitive U-SVM for different types of Universa (digit "1", "3", "6" and RA) and different cost-ratios (r = 0.5, 0.2, 0.1). Typical histograms of projections for training data along with the Universum are shown in Figs. 10 and 11. Note that histograms for the cost-ratio r=0.2 are not shown, because they are very similar to histograms for r=0.1.

The histograms of projections in Figs. 10-11 have the following characteristics:

- positive and negative training samples are well-separable.
- *digit* '1': *well spread* Universum *outside* training samples' class means and *highly biased* towards positive class.
- *digit* '3': *well spread* Universum samples about training samples' class means and *slightly biased* towards negative class.
- *digit* '6': *well spread* Universum samples about training samples class means but *slightly biased* towards positive class.
- *Random Averaging*: *well spread* universum samples about training samples' class means but *slightly biased* towards positive class.

Practical conditions (B1)-(B3) indicate that for the given well-separable training samples (digit '8' vs. '5'); digit '3' is the best choice for Universum. Although, digit '6' and RA are well-spread about the training samples' class means; they are slightly biased towards positive class. Further, digit '1' samples represent the worst choice, as they are not well-spread about the training samples' class means, and are highly biased towards positive class. These findings are consistent with empirical results in Table 6, showing no statistically meaningful improvement for digit 1 Universum, and a good improvement for digits '3', digit '6' and RA.

The *4th set of experiments* also involves the classification of handwritten digits "8" vs. "5" using MNIST data. We use the same experimental setup with 1000 training/validation samples, and introduce artificial Universum samples formed as follows. Each component (pixel) of a 28 × 28 = 784 dimensional sample follows a binomial distribution with probability p(x = 1) = 0.1395. This probability value 0.1395 is chosen so that the average intensity of the Universum samples is the same as that of the training data (averaged for both digits 5 and 8). Fig. 12(a) shows an example of such a universum. Intuitively, this (*random noise*) Universum is not expected to improve the generalization of cost-sensitive SVM.

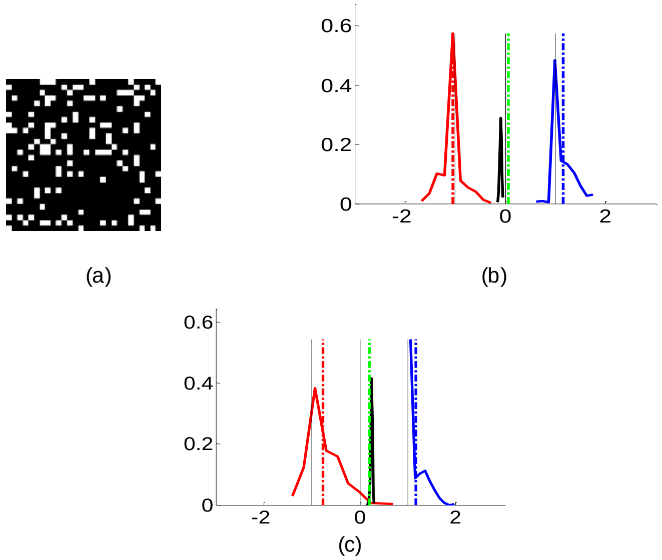


Fig. 12. Binomially distributed Universum (random noise).
(a) 28x28 image.
(b) Histogram of projections for cost-ratio r=0.5 ($C^*/C=2^{-15}$).
(c) Histogram of projections for cost-ratio r=0.1 ($C^*/C=2^{-15}$).

TABLE 7. COMPARISON OF COST-SENSITIVE SVM AND COST-SENSITIVE U-SVM WITH BINOMIALLY DISTRIBUTED UNIVERSUM FOR DIFFERENT COST-RATIOS

METHODS	cost-sensitive SVM	cost-sensitive U-SVM (RA)
	Cost-Ratio (r=0.5)	
test error (in %)	**1.46(0.32)**	**1.46(0.32)**
FP rate (in %)	1.19(0.30)	1.19(0.30)
FN rate (in %)	1.58(0.54)	1.58(0.54)
	Cost-Ratio (r=0.2)	
test error (in %)	**1.36(0.36)**	**1.35(0.35)**
FP rate (in %)	3.61(2.37)	3.59(2.33)
FN rate (in %)	0.95(0.37)	0.95(0.37)
	Cost-Ratio (r=0.1)	
test error (in %)	**1.16(0.07)**	**1.11(0.11)**
FP rate (in %)	7.98(4.13)	7.17(3.94)
FN rate (in %)	0.53(0.41)	0.55(0.39)

Experimental results comparing the test error rates for cost-sensitive RBF SVM classifier and U-SVM using 1,000 Universum samples are shown in Table 7. The histograms of projections are provided in Figs. 12(b)-(c). As expected, this Universum does not yield any improvement (over cost-sensitive SVM). This can be anticipated from the histogram of projections in Fig. 12, because projections of the Universum samples are not well-spread about the class means.

The *5th set of experiments* uses the *Real-life ISOLET data set* [22], 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 [22]. 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:

- Number of training samples ~ 100 (50 samples per class).
- Number of Universum samples ~ 300 (three types of Universa: letters D, P and RA).
- Number of validation/test samples ~ 500 (250 samples per class).

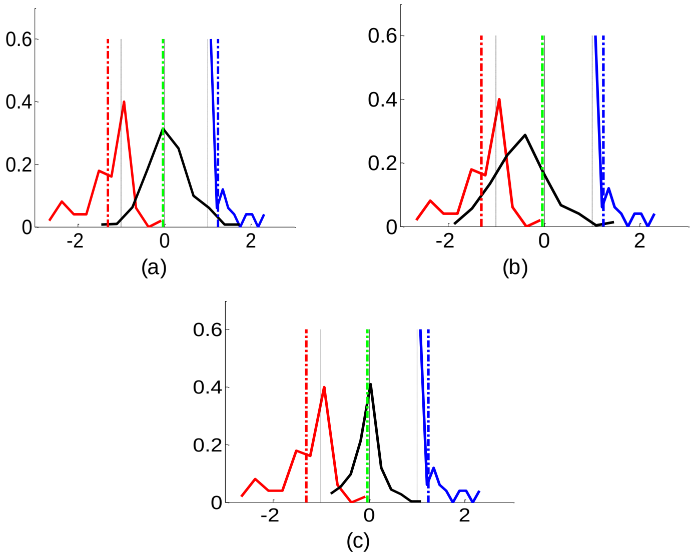


Fig. 13. Univariate histograms of projections for cost-sensitive SVM with r=0.5 (C=2^{-4}) for different types of Universa. Training set size ~100 samples. Universum set size ~300 samples. (a) letter D Universum C*/C=2^{-4} (b) letter P Universum C*/C=2^{-5} (c) RA Universum C*/C=2^{-4}.

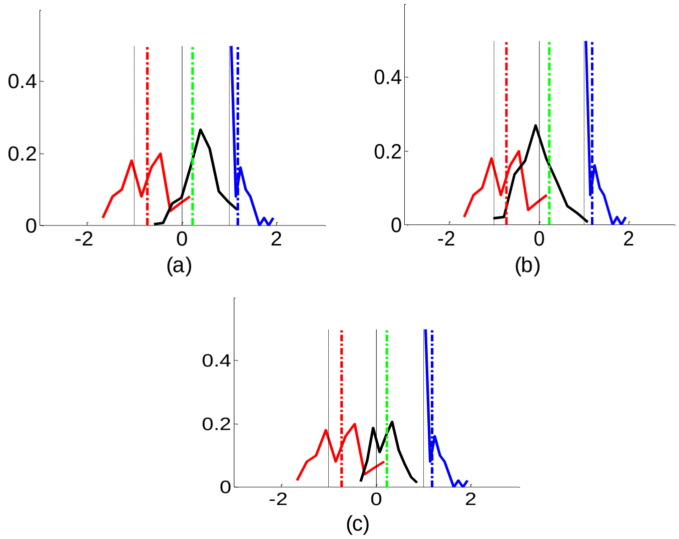


Fig. 14. Univariate histogram of projections for cost-sensitive SVM with r=0.1 (C=2^{-3}) for different types of Universa. Training set size ~100 samples. Universum set size ~300 samples. (a) letter D Universum C*/C=2^{-10} (b) letter P Universum C*/C=2^{-6} (c) RA Universum C*/C=2^{-5}.

TABLE 8. COMPARISON OF STANDARD SVM, COST-SENSITIVE SVM AND COST-SENSITIVE U-SVM ON ISOLET ('B' VS. 'V' DATASET) FOR DIFFERENT COST-RATIOS

METHODS	standard SVM	cost-sensitive SVM	cost-sensitive U-SVM (letter D)	cost-sensitive U-SVM (letter P)	cost-sensitive U-SVM (RA)
			Cost-Ratio (r=0.5)		
test error (in %)	**5.34(1.47)**	**5.21(1.23)**	**4.59(1.24)**	**4.33(0.82)**	**4.96(1.05)**
FP rate (in %)	9.36(2.31)	10.20(4.08)	10.32(3.88)	10.20(3.78)	9.52(3.40)
FN rate (in %)	3.32(1.61)	2.72(1.90)	1.72(1.21)	1.40(0.84)	2.68(1.85)
			Cost-Ratio (r=0.2)		
test error (in %)	**3.51(0.51)**	**3.42(0.42)**	**2.93(0.61)**	**2.77(0.52)**	**3.03(0.53)**
FP rate (in %)	11.68(3.20)	12.56(3.18)	12.6(3.83)	13.6(3.76)	11.96(2.59)
FN rate (in %)	1.88(0.98)	1.60(0.75)	1.00(0.74)	0.60(0.43)	1.24(0.74)
			Cost-Ratio (r=0.1)		
test error (in %)	**2.79(0.75)**	**2.70(0.65)**	**2.59(0.58)**	**1.78(0.42)**	**2.39(0.69)**
FP rate (in %)	12.24(3.81)	15.28(4.28)	14.88(4.07)	17.6(4.82)	14.6(3.93)
FN rate (in %)	1.84(0.76)	1.44(0.66)	1.36(0.60)	0.2(0.28)	0.48(0.45)

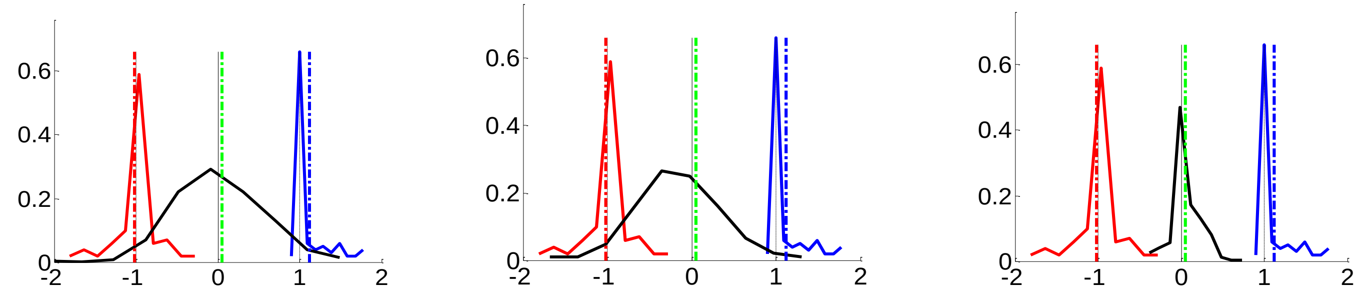
Our initial experiments suggest that linear SVM works well for this dataset. Comparisons of the (linear) standard SVM, cost-sensitive SVM and the cost-sensitive U-SVM for the different types of Universa: letters D, P and RA with different cost-ratios (r=0.5, 0.2, 0.1) are shown in Table 8. Typical histograms of projections for training data along with the Universum data for the cost-ratios (r=0.5, 0.1) are shown in Figs 13 and 14. For this dataset, typical histograms of projections for the cost-ratio r=0.2 are very similar to r=0.1, and have been omitted. 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 are *slightly biased* towards the negative class.
- *letter D* has *narrower* projections than *letter P* and they are *slightly biased* towards the positive class.
- *Random Averaging* has *narrower* projections than *letter P* and they are *slightly biased* towards positive class.

Hence, based on conditions (B1)-(B3), *letter P* is expected to be more effective than *letter D* and RA. This is consistent with empirical results in Table 8.

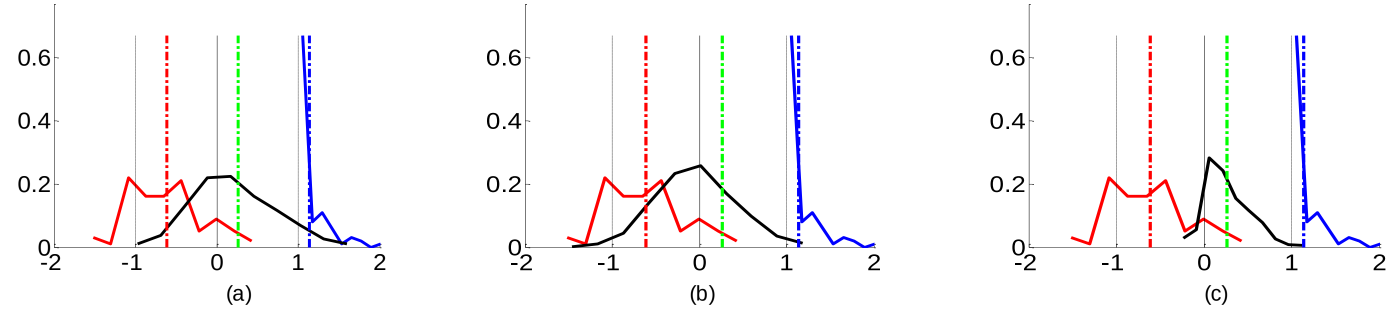
TABLE 9. COMPARISON OF STANDARD SVM, COST-SENSITIVE SVM AND COST-SENSITIVE U-SVM ON GTSRB ('50' VS. '80' DATASET) FOR DIFFERENT COST-RATIOS

METHODS	standard SVM	cost-sensitive SVM	cost-sensitive U-SVM (sign 30)	cost-sensitive U-SVM (sign 60)	cost-sensitive U-SVM (RA)
	Cost-Ratio (r=0.5)				
test error (in %)	**9.82(0.83)**	**9.25(0.99)**	**6.75(1.09)**	**6.84(1.30)**	**8.91(0.60)**
FP rate (in %)	6.74(1.56)	8.84(3.77)	8.98(4.81)	9.78(5.53)	8.20(2.91)
FN rate (in %)	11.36(1.74)	9.46(1.65)	5.64(2.49)	5.38(1.73)	9.26(1.03)
	Cost-Ratio (r=0.2)				
test error (in %)	**9.27(1.25)**	**7.14(1.26)**	**5.88(0.82)**	**5.93(0.98)**	**6.91(1.13)**
FP rate (in %)	7.12(1.79)	19.75(4.97)	23.8(5.54)	26.07(3.57)	16.25(4.70)
FN rate (in %)	9.7(1.6)	4.63(2.06)	2.3(1.51)	1.9(0.91)	5.05(2.06)
	Cost-Ratio (r=0.1)				
test error (in %)	**9.44(1.70)**	**5.71(1.05)**	**4.74(1.15)**	**4.62(1.28)**	**4.77(0.75)**
FP rate (in %)	6.64(2.19)	45.02(18.69)	42.54(14.16)	44.98(14.27)	26.68(7.33)
FN rate (in %)	9.72(1.98)	1.78(1.32)	0.96(0.49)	0.58(0.45)	2.6(1.00)



(c)

Fig. 15. Univariate histogram of projections for cost-sensitive SVM with cost-ratio r=0.5 ($C=2^{-2}$) for different types of Universa. Training set size ~200 samples. Universum set size ~1000 samples. (a) sign '30' Universum $C^*/C=2^{-2}$ (c) sign '60' Universum $C^*/C=2^{-4}$ (c) RA Universum $C^*/C=2^{-5}$.

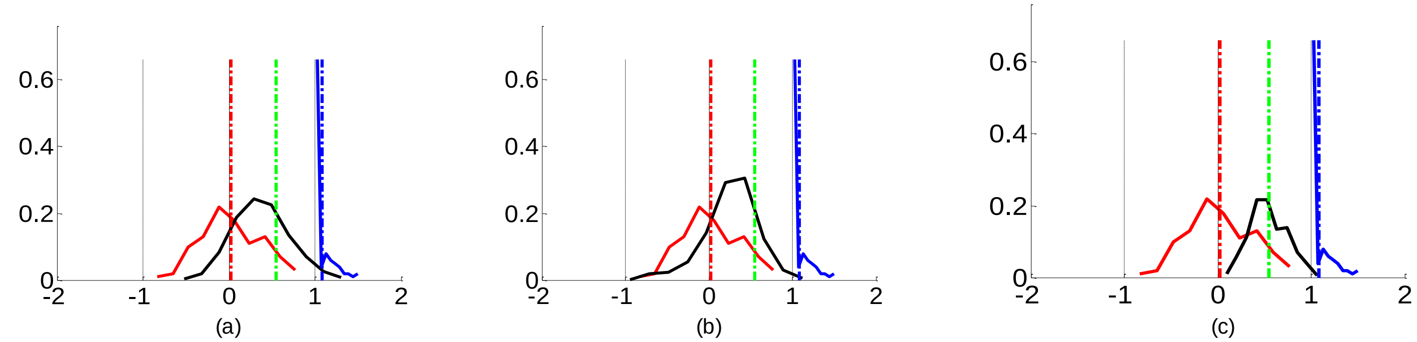


(a)

(b)

(c)

Fig. 16. Univariate histogram of projections for cost-sensitive SVM with cost-ratio r=0.2 ($C=2^{-2}$) for different types of Universa. Training set size ~200 samples. Universum set size ~1000 samples. (a) sign '30' Universum $C^*/C=2^{-4}$ (c) sign '60' Universum $C^*/C=2^{-4}$ (c) RA Universum $C^*/C=2^{-7}$.



(a)

(b)

(c)

Fig. 17. Univariate histogram of projections for cost-sensitive SVM with cost-ratio r=0.1 ($C=2^{-2}$) for different types of Universa. Training set size ~200 samples. Universum set size ~1000 samples. (a) sign '30' Universum $C^*/C=2^{-7}$ (c) sign '60' Universum $C^*/C=2^{-6}$ (c) RA Universum $C^*/C=2^{-6}$.

For our 6th *set of experiments* we use the real-life German Traffic Sign Recognition Benchmark (GTSRB) dataset [23]. The task is to perform traffic sign classification between the images of the signs "50" vs. "80". These sample images are represented by their pyramid histogram of gradient (PHOG) features [10, 23]. We label the traffic sign '50' as class '+1' and the traffic sign '80'as class '-1'. The cost-ratio is specified as,

$$r = \frac{C_{fp}}{C_{fn}} = \frac{\text{missclassification cost(truth='80',prediction='50')}}{\text{missclassification cost(truth='50',prediction='80')}}$$

For this experiment:

- Number of training samples ~200 (100 per class).
- Number of validation samples ~ 200 (100 per class).
- Number of Universum samples ~ 1000 (3 types of Universa: signs '30', '60' and RA).
- Number of Test samples ~ 2000 (1000 per class).
- Dimensionality of the input space ~ 1568 (PHOG features).

Initial experiments suggest that linear parameterization is optimal for this dataset; hence only linear kernel has been used in all comparisons. Performance comparisons between standard SVM, cost-sensitive SVM and cost-sensitive U-SVM for the different types of Universa: signs '30', '60' and RA with different cost-ratios (r=0.5, 0.2, 0.1) are shown in Table 9. The typical histograms of projections for training data along with the Universum data are also shown in Fig. 15, 16 and 17. Analysis of projections for different types of universum samples shows that:

- *sign* '30' has *well spread* projections between the training samples' class means and *slightly biased* towards the negatve class.
- *sign* '60' has *well spread* projections between the training samples' class means and *slightly biased* towards the negatve class.
- *Random Averaging* has *narrower* projections than projections for the signs "30" and "60", except for the cost-ratio r=0.1, for which it has well-spread projections about the training samples' class means.

Hence, for the cost-ratios r=0.5, 0.2 we can expect signs "30" and "60" to be more effective than RA Universum. Further, for r=0.1 all the three types of Universa are likely to provide similar improvements in generalization performance. This is consistent with the empirical results in Table 9.

Our final experiment uses publicly available Freiburg Electroencephalogram (EEG) dataset [24]. The dataset contains intracranial EEG recordings from 21 patients with medically intractable focal epilepsy. For each patient, the dataset contains EEG recordings from 6 electrodes, sampled at 256 samples/sec. These EEG signals have been labeled by human medical experts as preictal (30 min preceding a seizure onset), ictal and interictal, as shown in Fig. 18. The goal is to estimate predictive model for discriminating between preictal and interictal signals. This model should be estimated from the training data with known class labels. This can be formalized as a binary classification problem.

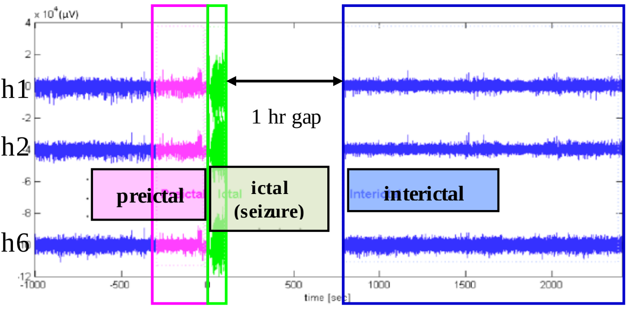
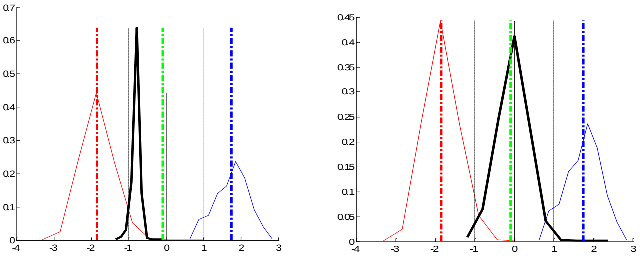


Fig. 18. iEEG recordings from six electrodes with a seizure event (ictal, shown in green) reproduced from [25]. Preictal signals (30 min preceding a seizure onset) are shown in pink. Interictal signals (at least 1 hour preceding or postceding a seizure) are shown in blue.



(a) (b)

Fig. 19. Univariate histogram of projections onto cost-sensitive SVM normal weight vector (C=10, $\gamma =$ 0.25) for different types of Universa for r=0.1. (a) *patient_20* interictal with C*/C=2^{-5} (b) Random Averaging with C*/C= 4.

Unknown nature of epileptic seizures and high variability of EEG patterns across patients favor *patient-specific* predictive modeling. That is, a separate classifier is estimated for each patient in the Freiburg dataset (using labeled training data for this patient). In our experiments (reported below), the task is to classify 'preictal' vs. 'interictal' signals for *patient_1* in the Freiburg dataset. Further, available data is highly unbalanced because seizure events are quite rare: there are approximately 10 times more interictal samples than preictal in the Freiburg dataset. Cost-sensitive SVMs are used to account for unequal misclassification costs common in biomedical applications [15, 25, 26]. Our experiments use the cost-ratio specified as:

$$r = \frac{C_{fp}}{C_{fn}} = \frac{\text{(truth='interictal',prediction=preictal')}}{\text{(truth='preictal',prediction='interictal')}} = \frac{1}{10}$$

The input features used for SVM modeling have been generated using preprocessing and feature selection steps described in [25], as follows. As a part of pre-processing, standard bipolar and/or time-differential methods have been applied to remove/reduce noise in EEG signals [25, 26]. Then EEG signals were divided into 20 sec windows with a 10 sec overlap. For each window, the Power Spectral Density (PSD) of nine different spectral bands: delta (0.5-4 Hz), theta (4-8

TABLE 10. COMPARISON OF STANDARD SVM, COST-SENSITIVE SVM AND COST-SENSITIVE U-SVM ON FREIBURG DATASET FOR PATIENT 1. ('PREICTAL' VS. 'INTERICTAL')

METHODS	standard SVM	cost-sensitive SVM	cost-sensitive U-SVM (patient '20')	cost-sensitive U-SVM (RA)
test error (in %)	**15.74**	**12.69**	**12.18**	**5.16**
FP rate (in %)	0.09(2/2154)	0(0/2154)	0(0/2154)	0.09(2/2154)
FN rate (in %)	17.3(31/179)	13.9(25/179)	13.4(24/179)	5.5(10/179)

Hz), alpha (8-13 Hz), beta (13-30 Hz), four gamma bands (30-47 Hz, 53-70Hz , 70-90 Hz and 103-128 Hz) were computed for all the 6 electrodes. Each moving window is represented as an input feature vector of size 6 x 9 = 54, and each window in the training data is labeled as interictal (negative) or preictal (positive). These 54-dimensional training samples are used to estimate an SVM classifier, in order to predict future (unlabeled) test inputs. For this experiment, available data contains 4 seizure recordings for the *patient_1*. Hence, seizures 1, 2, and 3 are used for training and seizure 4 is used as test data. Our goal is to investigate the effectiveness of the cost-sensitive Universum SVM for modeling *patient-1* data. To this end, we used two different types of Universa: interictal signals of other patients, and random averaging (RA). All Universum modeling results using interictal data from other patients showed similar (poor) performance, so we present (below) results only for the Universum formed using *patient_20* interictal data. A brief description of the experimental setting is provided below:

- Number of training samples (seizures '1','2' and '3') ~ 6999 ('preictal' ~ 537 and 'interictal' ~ 6462 samples).
- Number of Universum samples ~ 7000 (2 types of Universa: *patient_20* interictal and RA).
- Number of test samples (seizure '4') ~ 2333 (preictal' ~ 179 and 'interictal' ~ 2154 samples)
- Dimensionality of each sample = 54 (9 spectral bands × 6 electrodes).

Following [25], we use an RBF kernel of the form $K(\mathbf{x},\mathbf{x}') = \exp\left(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2\right)$. Further, SVM model selection is performed via 5-Fold cross-validation procedure on the training data. Performance comparisons between standard SVM, cost-sensitive SVM and cost-sensitive U-SVM for two different types of Universa are shown in Table 10. For all the methods we have training error ~ 0%. Typical histograms of projections for training data along with the Universum data are also shown in Fig. 19. Visual analysis of these histograms of projections indicates that:

- *patient '20' interictal* has very *narrow* projections between the training samples' class means .
- *Random Averaging* has *well spread* projections between the training samples' class means.

Hence, we can expect the RA Universum to be more effective than *patient '20' interictal*. This is consistent with the empirical results in Table 10.

5 CONCLUSIONS

Previous studies [2-10] have demonstrated the effectiveness of the Universum learning for improving the generalization of SVM classifiers. However, all these studies used balanced data sets with equal misclassification costs. This paper describes new U-SVM formulation that incorporates different misclassification costs and can be used for unbalanced data sets. The proposed cost-sensitive U-SVM can be implemented using minor modifications to existing U-SVM software. This modified software is made publicly available at [18].

We also presented practical conditions for the effectiveness of the cost-sensitive U-SVM using analysis of the histograms of projections. These proposed conditions also hold for unbalanced data sets typically seen in many biomedical/bioinformatics applications. These conditions can be adopted by practitioners 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 the cost-sensitive U-SVM to that of cost-sensitive SVM.

According to our analysis, meaningful characterization of 'good' Universum is possible only in the context of a particular labeled training dataset. This point is particularly important for biomedical applications, where predictive data-analytic models are often patient-specific (as in the seizure prediction example in Section 4). In these applications, there is no good medical/clinical intuition about good Universa. Hence, the proposed conditions (for the effectiveness of the Universum learning under cost-sensitive settings) are expected to be quite useful.

Finally, we point out that many applications involve extreme scenarios with very high cost ratios or extreme unbalance in the data (as in anomaly detection). Such problems follow a different learning framework called single-class learning [11, 15], and has not been explored in this work. Hence, there is a need for future research on the effectiveness of Universum learning under such extreme settings.

ACKNOWLEDGMENT

This work was supported, in part, by NSF grant ECCS-0802056. The authors gratefully acknowledge multiple discussions with Prof. Vladimir Vapnik from Columbia University on the Universum SVM learning. We also gratefully acknowledge help from Dr. Yun Park who provided preprocessed EEG data originally used in [25].

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, and H. Wang, "UBoost: Boosting with the Universum", IEEE Transaction on Pattern Analysis and Machine Intelligence, 2011.

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

[12] G. M. Weiss, K. McCarthy, and 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] S. Dhar and V. Cherkassky, "Cost-Sensitive Universum-SVM", *ICMLA* 2012.

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

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

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

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

[19] 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.

[20] F. Cai, V. Cherkassky, D. Weisdorf, M. Arora, B. Van Ness, "Predictive modeling of Transplant-Related Mortality," *Proc. of the 2010 Design of Medical Devices Conf.*, Minneapolis, April 2010.

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

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

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

[24] Seizure Prediction Project Freiburg. [WWW page]. URL https://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database

[25] Y. Park, T. Netoff, and K. Parhi, "Seizure Prediction with Spectral Power of Time/Space-Differential EEG Signals using Cost-Sensitive Support Vector Machine",ICASSP, Dallas,TX, March 2010, pp. 5450-5453.

[26] Y. Park, L. Luo, K. Parhi, and T. Netoff, "Seizure Prediction with Spectral Power of EEG using Cost-Sensitive Support Vector Machines", *Epilepsia*, 52(10), pp. 1761-1770, Wiley 2011.