On Multiclass Universum Learning

Sauptik Dhar[†] **Naveen Ramakrishnan**[†] **Vladimir Cherkassky**[‡] **Mohak Shah**[†*]
[†]Bosch Research Center, CA [‡] University of Minnesota, MN [*] University of Illinois at Chicago, IL
{sauptik.dhar, naveen.ramakrishnan, mohak.shah}@us.bosch.com cherk001@umn.edu

Introduction 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. Such high-dimensional data sets present new challenges for most learning problems. Recent studies have shown Universum learning to be particularly effective for such high-dimensional low sample size data settings [1–11]. However, most such studies are limited to binary classification problems. This paper introduces universum learning for multiclass SVM [12] under balanced settings with equal misclassification costs and propose a new formulation called multiclass Universum SVM (MU-SVM). We provide empirical results in support of the proposed formulation.

Universum Learning for Multiclass SVM The idea of Universum learning was introduced by Vapnik for binary classification problems [13, 14] to incorporate a priori knowledge about admissible data samples. Here, in addition to labeled training data we are also given a set of unlabeled examples from the Universum which belongs to the same application domain as the training data but are known not to belong to either class. In fact, this idea can also be extended to multiclass problems. For multiclass problems in addition to the labeled training data we are also given a set of unlabeled Universum examples. However, now the Universum samples are known not to belong to *any* of the classes in the training data. For example, if the goal of learning is to discriminate between handwritten digits 0, 1, 2,...,9; one can introduce additional 'knowledge' in the form of handwritten letters A, B, C, ... ,Z. These examples from the Universum contain certain information about handwriting styles, but they cannot be assigned to any of the classes (1 to 9). These Universum samples are introduced into the learning as contradictions and hence should lie close to the decision boundaries for all the classes $\mathbf{f} = [f_1, \ldots, f_L]$. This argument follows from [14, 15], where the universum samples lying close to the decision boundaries are more likely to falsify the classifier. To ensure this, we incorporate a Δ - insensitive loss function for the universum samples which forces the universum samples to lie close to the decision boundaries ('0' in Fig. 1) for all the classes i.e. $\mathbf{f} = [f_1, \ldots, f_L]$. This reasoning motivates the new multiclass Universum-SVM (MU-SVM) formulation where: 1) Standard hinge loss is used for the training samples [12] 2) The universum samples are penalized by a Δ - insensitive loss (Fig. 1) for the decision functions of all the classes $\mathbf{f} = [f_1, \ldots, f_L]$. This leads to the following MU-SVM formulation. Given training samples $\mathcal{T} := (\mathbf{x}_i, y_i)_{i=1}^n$, where $y_i \in \{1, \ldots, L\}$ and additional unlabeled universum samples $\mathcal{U} := (\mathbf{x}_j^*)_{j=1}^m$. Solve [1],

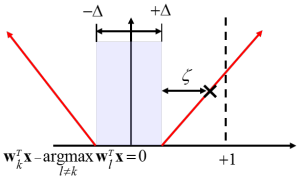

Figure 1: Loss function for universum samples for k^{th} decision function $f_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x}$. An universum sample lying outside the Δ- insensitive zone is penalized linearly using the slack variable ζ.

$$\min_{\mathbf{w}_1 \ldots \mathbf{w}_L, \xi, \zeta} \quad \frac{1}{2}\sum_l \|\mathbf{w}_l\|_2^2 \quad + \quad C\sum_{i=1}^n \xi_i \quad + \quad C^*\sum_{j=1}^m \zeta_j \tag{1}$$

$$s.t. \quad (\mathbf{w}_{y_i} - \mathbf{w}_l)^\top \mathbf{x}_i \geq e_{il} - \xi_i; \quad e_{il} = 1 - \delta_{il}, \quad i = 1 \ldots n$$

$$|(\mathbf{w}_k - \mathbf{w}_l)^\top \mathbf{x}_j^*| \leq \Delta + \zeta_j; \quad j = 1 \ldots m, \quad l, k = 1 \ldots L$$

Here, the universum samples that lie outside the Δ - insensitive zone are linearly penalized using the slack variables $\zeta_j \geq 0, j = 1 \ldots m$. The user-defined parameters $C, C^* \geq 0$ control the trade-off

[1] Throughout this paper, we use index i for training samples, j for universum samples and k, l for the class labels.

Figure 2: Traffic Signs. (a) '30' (b) '70' (c) '80' (d) 'no-entry'.

Table 1: SVM vs. MU-SVM. Mean test error in %, over 10 runs. (std. deviation in parenthesis).

SVM	**MU-SVM** (*'no-entry'*)
7.47(0.92)	6.57(0.59)

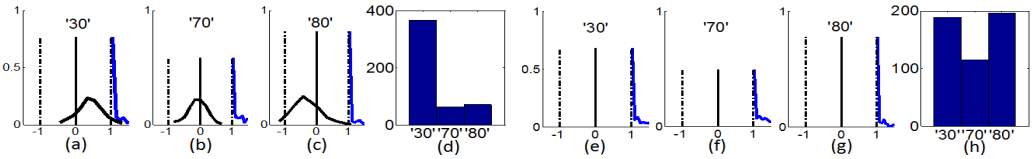

Figure 3: Typical histogram of projection of training samples (shown in **blue**) and universum samples (shown in **black**). SVM decision functions for (a) sign '30'. (b) sign '70'.(c) sign '80'. (d) frequency plot of predicted labels for universum samples for estimated SVM model. U-SVM decision functions for (e) sign '30'. (f) sign '70'.(g) sign '80'. (h) frequency plot of predicted labels for universum samples for estimated MU-SVM model.

between the margin size, the error on training samples, and the contradictions (samples lying outside $\pm\Delta$ zone) on the universum samples.

Results For our empirical results we use *German Traffic Sign Recognition Benchmark (GTSRB) dataset* [16]. The goal here is to identify the traffic signs '30','70' and '80' (Fig. 2(a)-(c)) represented by their histogram of gradient (HOG) features ($\sim$ 1568 dimensions). Further, in addition to the training samples we are also provided with additional universum samples i.e. traffic signs for '*no-entry*' (see Fig. 2d). For this experiment we use the following setting,
– No. of training/test samples = 300 (100 per class)/1500 (500 per class) respectively.
– No. of universum samples = 500 (additional samples did not improve performance).
Initial experiments suggest that linear parameterization is optimal for this dataset. Here, the model selection is done over the range of parameters, $C = [10^{-4}, \ldots, 10^{3}]$, $C^*/C = \frac{n}{mL} = 0.2$ and $\Delta = [0, 0.01, 0.05, 0.1]$ using stratified 5-Fold cross validation [17]. Performance comparisons between SVM and MU-SVM are shown in Table 1. Table 1 shows that the MU-SVM model provides better generalization than the multiclass SVM model. For better understanding of the MU-SVM modeling results we adopt the technique of '*histogram of projections*' originally introduced for binary classification in [18, 19]. However, different from binary classification, here we project the training samples onto the decision space for that class; and the universum samples onto the decision spaces of all the classes. Additionally, we also generate the frequency plot of the predicted labels for the universum samples. Fig 3 shows the typical histograms and frequency plots for the SVM and MU-SVM models using the '*no-entry*' sign (as universum). As seen from Fig. 3(a)-(c), the optimal SVM model has high separability for the training samples i.e., most of the training samples lie outside the margin borders with training error ~ 0. Infact, similar to binary SVM [19], we see data-piling effects for the training samples near the '+1' - margin borders of the decision functions. This is typically seen for high-dimensional low sample size settings. However, the universum samples are widely spread about the margin-borders. Moreover, the universum samples are biased towards the positive side of the decision boundary of the sign '30' (see Fig 3(a)) and hence predominantly gets classified as sign '30'(see Fig.3 (d)). As seen from Figs 3 (e)-(g), applying the MU-SVM model preserves the separability of the training samples and additionally reduces the spread of the universum samples. For such a model the uncertainity due to universum samples is uniform across all the classes i.e. signs '30','70' and '80' (see Fig. 3(h)). The resulting MU-SVM model has higher contradiction on the universum samples and provides better generalization in comparison to SVM. Additional results and analysis are available in [20].

Conclusion The results show that the proposed MU-SVM provides better performance than multiclass SVM, typically for high-dimensional low sample size settings. Under such settings the training data exhibits large data-piling effects near the margin border ('+1'). For such ill-posed settings, introducing the Universum can provide improved generalization over the multiclass SVM solution. However, the effectiveness of the MU-SVM also depends on the properties of the universum data. Such statistical characteristics of the training and universum samples for the effectiveness of MU-SVM can be conveniently captured using the 'histogram-of-projections' method introduced in this paper. This is open for future research.

References

[1] F. Sinz, O. Chapelle, A. Agarwal, and B. Schölkopf, "An analysis of inference with the universum," in *Advances in neural information processing systems 20*. NY, USA: Curran, Sep. 2008, pp. 1369–1376.

[2] S. Chen and C. Zhang, "Selecting informative universum sample for semi-supervised learning." in *IJCAI*, 2009, pp. 1016–1021.

[3] S. Dhar and V. Cherkassky, "Development and evaluation of cost-sensitive universum-svm," *Cybernetics, IEEE Transactions on*, vol. 45, no. 4, pp. 806–818, 2015.

[4] S. Lu and L. Tong, "Weighted twin support vector machine with universum," *Advances in Computer Science: an International Journal*, vol. 3, no. 2, pp. 17–23, 2014.

[5] Z. Qi, Y. Tian, and Y. Shi, "A nonparallel support vector machine for a classification problem with universum learning," *Journal of Computational and Applied Mathematics*, vol. 263, pp. 288–298, 2014.

[6] C. Shen, P. Wang, F. Shen, and H. Wang, "Uboost: Boosting with the universum," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 825–832, 2012.

[7] Z. Wang, Y. Zhu, W. Liu, Z. Chen, and D. Gao, "Multi-view learning with universum," *Knowledge-Based Systems*, vol. 70, pp. 376–391, 2014.

[8] D. Zhang, J. Wang, F. Wang, and C. Zhang, "Semi-supervised classification with universum." in *SDM*. SIAM, 2008, pp. 323–333.

[9] Y. Xu, M. Chen, and G. Li, "Least squares twin support vector machine with universum data for classification," *International Journal of Systems Science*, pp. 1–9, 2015.

[10] Y. Xu, M. Chen, Z. Yang, and G. Li, "ν-twin support vector machine with universum data for classification," *Applied Intelligence*, vol. 44, no. 4, pp. 956–968, 2016.

[11] C. Zhu, "Improved multi-kernel classification machine with nyström approximation technique and universum data," *Neurocomputing*, vol. 175, pp. 610–634, 2016.

[12] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," *Machine learning*, vol. 47, no. 2-3, pp. 201–233, 2002.

[13] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[14] V. Vapnik, *Estimation of Dependences Based on Empirical Data (Information Science and Statistics)*. Springer, Mar. 2006.

[15] J. Weston, R. Collobert, F. Sinz, L. Bottou, and V. Vapnik, "Inference with the universum," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 1009–1016.

[16] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, pp. –, 2012.

[17] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[18] V. Cherkassky, S. Dhar, and W. Dai, "Practical conditions for effectiveness of the universum learning," *Neural Networks, IEEE Transactions on*, vol. 22, no. 8, pp. 1241–1255, 2011.

[19] V. Cherkassky and S. Dhar, "Simple method for interpretation of high-dimensional nonlinear svm classification models." in *DMIN*, R. Stahlbock, S. F. Crone, M. Abou-Nasr, H. R. Arabnia, N. Kourentzes, P. Lenca, W.-M. Lippe, and G. M. Weiss, Eds. CSREA Press, 2010, pp. 267–272.

[20] S. Dhar, N. Ramakrishnan, V. Cherkassky, and M. Shah, "Universum learning for multiclass svm," *arXiv preprint arXiv:1609.09162*, 2016.