Simple Method for Interpretation of High-Dimensional Nonlinear SVM Classification Models

Vladimir Cherkassky, Fellow, IEEE and Sauptik Dhar
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis MN 55455 USA

Abstract - *Many applications in machine learning involve modeling sparse high dimensional data. Examples include application of predictive models in high dimensional micro-array data, or in brain imaging studies using magnetic resonance imaging (MRI). A typical problem in such a setting is the understanding of the multivariate models estimated from the data, especially nonlinear high-dimensional models such as Support Vector Machines (SVM). In this paper we present a simple graphical method for interpreting such high-dimensional models and illustrate its effectiveness for improved understanding of SVM models and for analyzing new learning setting called Universum SVM.*

Keywords: Interpretation of black box models, model selection, support vector machines, univariate histogram of projections, Universum learning.

1 Introduction

Modeling sparse high dimensional data is common in machine learning applications, and, in particular, in biomedical applications. For example, in micro-array data or brain imaging studies using magnetic resonance imaging (MRI) where the dimension (d) of each samples is quite large in comparison to the number of samples (n). Such high-dimensional settings pose new challenges for classification methods. Most learning methods developed in statistics, machine learning and data mining, such as decision trees, MARS, discriminant analysis, support vector machines, and AdaBoost, follow standard inductive learning problem setting [1-3].These techniques have been successfully used in many real-life applications [4]. Another approach to handling ill-posed high-dimensional classification problems is to adopt new non-standard learning formulations that incorporate a priori knowledge about application data and/or the goal of learning directly into the problem formulation (see [1], [5]). Examples include:

- Transduction. [5, 6]
- Inference through Contradictions. [5]
- Learning with Structured Data. [5]
- Multi-task Learning. ([7-9]).

This work was supported by NSF grant ECCS-0802056 and by the BICB grant from the University of Minnesota, Rochester.

These new learning settings reflect properties of real-life applications, and can result in improved generalization. However their acceptance by practitioners is hindered by their poor interpretation capability. Thus a simple graphical representation for such multivariate high-dimensional models is essential for their understanding and acceptance.

Recent approaches aimed at improved understanding of SVM classifiers include a number of standard visualization and graphical techniques originally developed in statistics and later adapted for SVM interpretation [10-12]. These methods follow a traditional statistical approach of identifying a few 'important' low-dimensional projections, and this may place a heavy burden on a human modeler trying to examine large number of possible projections. In many cases these methods are limited to linear kernels ([10, 11]) or put additional constraints upon the type of kernels used [12]. Further, these visualization methods are general-purpose, and they do not utilize many critical aspects of SVM classifiers, such as the soft-margins. Much can be understood from the visualization of how the data is oriented w.r.t. the soft-margins. In this paper we present a very simple graphical method called the "univariate histogram of projections" which can be used for many SVM-based methods, and demonstrate its practical utility for several representative data sets.

The paper is organized as follows. Section 2 provides a brief description of the univariate histogram of projections for both linear and non-linear SVM models. Section 3 describes how the proposed method can be used for SVM models with unbalanced data and unequal misclassification costs. Section 4 describes the utility of univariate histogram of projections for analyzing new SVM-based methodology called Universum Learning, or Inference through Contradiction [5]. Finally, the summary is presented in Section 5.

2 Univariate histogram of Projections

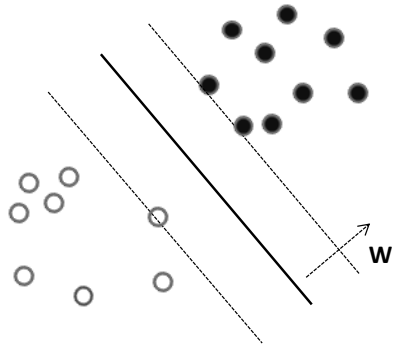
Classification methods specify (nonlinear) decision boundary in the space of input variables. This decision boundary is estimated from available training data, but is intended for classifying future (or test) input samples. For high-dimensional data, understanding and interpretation of both the training data and the estimated decision boundary is challenging, because (a) human intuition fails for such settings, and (b) high-dimensional sparse data sets have properties that are very different from low-dimensional settings [1].So we propose simple graphical representation of

the training data and SVM decision boundary (estimated from this data) via the "univariate histogram of projections".

Univariate Histogram of Projections ~ is the histogram of the projection values of the data samples onto the normal direction (weight vector) of the SVM decision boundary.

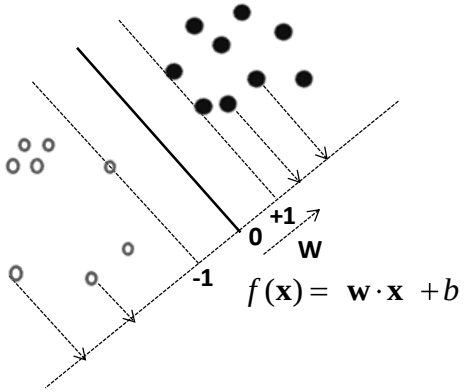
Such a histogram is obtained via the following three steps:-

a. Estimate standard SVM classifier for a given (labeled) training data set. Note that this step includes optimal model selection, i.e. tuning of SVM parameters (regularization parameter, kernel); (see Fig 1a).
b. Generate low-dimensional representation of training data by projecting it onto the normal direction vector of the SVM hyperplane estimated in (a); (see Fig 1b).
c. Generate the histogram of the projected values obtained in (b). (see Fig 1c).

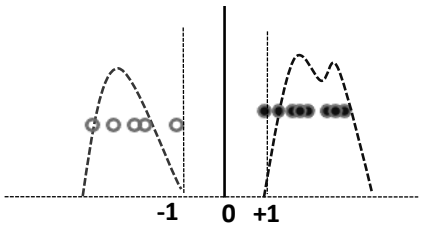


$$y = sign(f(\mathbf{x})) = sign(\mathbf{w}\cdot\mathbf{x} + b)$$

(a) The estimated SVM model and training data.



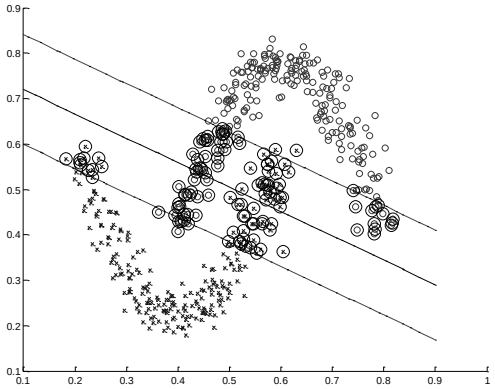
(b) Projection of the training data onto the normal weight vector (**w**) of the SVM hyperplane.



(c) Univariate histogram of Projections. i.e. histogram of $f(\mathbf{x})$ values.

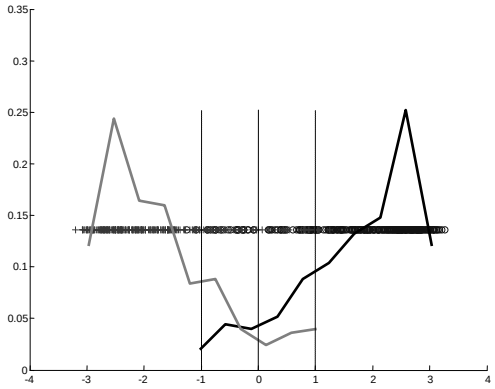
Fig 1. Illustration of the steps to generate the univariate histogram of projections.

It may be noted that the idea of projecting the data onto the normal direction given by **w**, is very similar to Fisher's Linear Discriminant Analysis, except that in this case the decision boundary is derived via nonlinear SVM. In Fig. 1c, SVM decision boundary is marked as zero, and the margin borders for positive/negative classes are marked, respectively, as +1/-1. Visual analysis of this univariate histogram of projections can be helpful for understanding high-dimensional data. For example, consider linear decision boundary for the synthetic Noisy Hyperbolas data set in Fig 2a. The projected values shown graphically in Fig. 2b are calculated analytically using linear SVM model as shown in Fig 2a and then the histogram of the projection of training samples onto the normal weight vector (**w**) are generated as shown in Fig 2b. The projected values for the two classes overlap, and this is consistent with the fact that the training samples are not linearly separable as seen from Fig 2a.



(a) Decision boundary for linear SVM model.

$$y = sign(f(\mathbf{x})) = sign(\mathbf{w}\cdot\mathbf{x} + b)$$

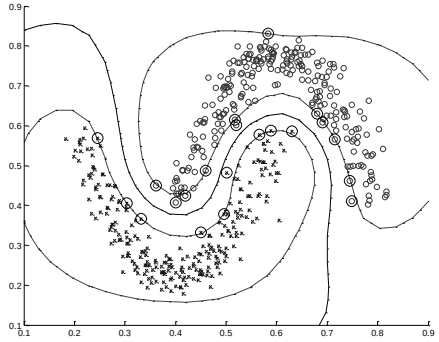


(b) Univariate Histogram of projections for training samples $\mathbf{x}_k$ onto the normal vector of linear SVM decision boundary.

$$f(\mathbf{x}_k) = (\mathbf{x}_k \cdot \mathbf{w}) + b$$

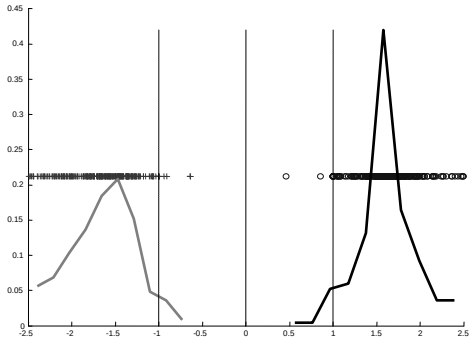
Fig 2. Example of the univariate histogram of projections for linear SVM.

For non-linear SVM kernels, the projected values $f(\mathbf{x})$ are calculated by using the kernel representation in the dual space. In this case, the projection of training sample $\mathbf{x}_k$ onto the normal direction of the nonlinear SVM decision boundary is expressed as $f(\mathbf{x}_k) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_k) + b$. For example, consider nonlinear decision boundary for the synthetic Noisy Hyperbolas data set in Fig 3a. Using nonlinear RBF kernel of the form $K(\mathbf{x}, \mathbf{x}') = \exp\ -\gamma \left\| \mathbf{x} - \mathbf{x}' \right\|^2$ with optimally tuned parameters yields SVM decision boundary shown in Fig 3a, and the corresponding histogram of projections in Fig 3b.



(a) Decision boundary for non-linear SVM model.

$$y = sign(f(\mathbf{x})) = sign(\sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b)$$



(b) Histogram of projections for training samples $\mathbf{x}_k$ onto the normal vector of non-linear SVM decision boundary.

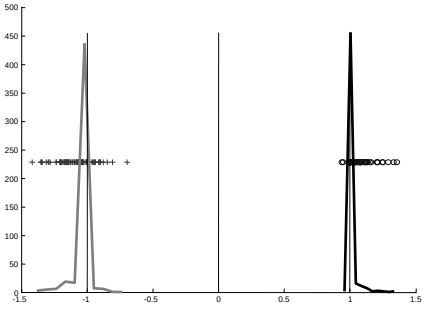
Fig.3.Example of the univariate histogram of projections for non linear RBF SVM.

Fig 3b clearly shows that the training samples are well separable using nonlinear decision boundary shown in Fig. 3a. In this case, the histogram does not add much to understanding, because separability of this two-dimensional data is evident from Fig. 3a. However, visual representation of high-dimensional data in the original input space, similar to Fig. 3a, is impossible, so the histogram representation becomes very useful.

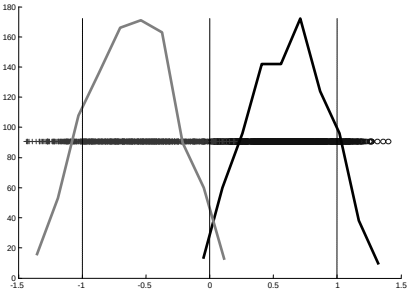
To illustrate this point, consider the sparse high dimensional *MNIST handwritten digit data set* [13], where training data samples represent handwritten digits 5 and 8, and the goal is to estimate binary classifier for discriminating these two digits. Each sample is represented as a real-valued vector of size 28*28=784. On average, 22% of the input features are non-zero which makes this data also sparse. For this example we use the following setting:-

- No. of training samples= 1000. (500 per class)
- No. of validation samples =1000. (This independent validation set is used for Model selection).
- No. of Test samples = 1866.
- Dimension of each sample= 784 (28*28).
- Range of parameters for RBF SVM model selection. C~[0.01, 0.1, 1, 10, 100, 100], and γ~ [2^{-8}, 2^{-6}… 2^2, 2^4].

For this data set, the univariate histogram of projection for the training samples as shown in Fig 4a. From Fig 4a we observe that the training samples are well separable in this optimally chosen RBF kernel space. This is typically the case for high dimension low sample size (HDLSS) setting, where the training samples are generally well separable in some optimally chosen kernel space. Of course, this property holds only for the training samples. The separability of the training samples does not imply separability for the test samples. This is illustrated in Fig 4b where the projections of test samples are not well separable.



(a)Histogram of projections of MNIST training data onto normal direction of RBF SVM decision boundary. Training set size ~ 1,000 samples.



(b) Histogram of projections of MNIST test data onto normal direction of RBF SVM decision boundary. Test set size ~ 1866 samples.

Fig.4. Univariate Histogram of projections for training/test samples for MNIST handwritten digits data set.

3 Histograms of projections for unbalanced data

Many practical applications use unbalanced data and different misclassification costs. This is common in most biomedical applications, fraud detection etc. Here 'unbalanced data' refers to the fact that the number of positive samples is much smaller than negative ones. Also, for such data sets, the cost of false negative errors C_{fn} is typically set higher than that of false positives C_{fp}. In this case, the ratio of misclassification costs is specified based on application-domain requirements [1].

In such a setting the quality of a classifier is estimated by its weighted classification error for test samples, with weights given by the misclassification costs, i.e.

$$weighted\_test\_error = C_{fp}P_{fp} + C_{fn}P_{fn} \qquad (1)$$

where P_{fp} and P_{fn} denote the probability of false positives and false negatives in the test set respectively [1]. In this case, the univariate histogram of projection can be quite helpful for interpreting the SVM models. Further such a simple graphical representation can also help the non expert SVM users (viz. clinicians and biomedical researchers) to interpret the black-box predictive SVM models, which in turn could make them acceptable in the medical research community.

As an example we consider a recent study of SVM predictive modeling of Transplant-Related Mortality (TRM) for Blood-and-Marrow Transplant patients [14]. In this study, the goal of modeling was to predict patient's survival (alive or dead) one year post-transplant. The data has records of 301 patients from a clinical study performed at the University of Minnesota. This data set has 221 samples labeled 'alive' and 75 samples labeled 'dead', i.e. the ratio of alive-to-dead samples is 3:1. Further, ratio of misclassification costs used in the study was 1:4 or 1:3. The prediction accuracy of classifier is measured using the weighted test error (1).

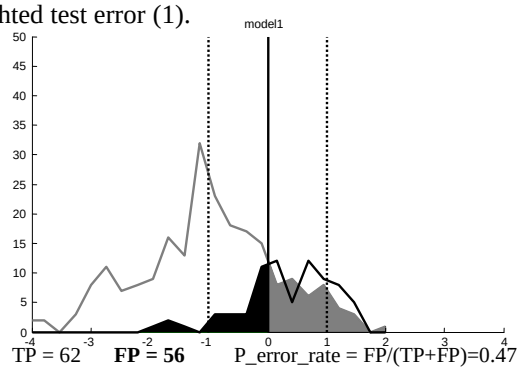


Fig. 5. Univariate histogram of projections for high-dimensional data. Negative samples (~ Alive) are shown in gray, positive (~Dead) in black. False positives and false negative errors are indicated as gray-shaded and black-shaded areas, respectively

Here we reproduce (in Fig. 5) the univariate histogram of projections for 14-dimensional SVM model for predicting TRM [14]. This figure illustrates unbalanced class distributions, and the effect of unequal misclassification costs on the performance of SVM classifier. In particular, as shown in Fig. 5, the error rate of patients classified as negative (Alive) is small (~7%), whereas the error rate of patients classified as positive (Dead) is large (~47%). Such a dichotomy is common for many practical problems with unbalanced data sets. Thus the proposed technique of univariate projections can also enable improved understanding of the bias of the SVM models for unbalanced data with unequal costs.

4 Conditions for effectiveness of Universum Learning

Sparse high-dimensional setting poses new challenges for classification methods. To this end we have seen the emergence of several non-standard learning formulations. One such learning method is 'inference through contradictions' or Universum learning [5, 6]. This idea was mainly introduced to incorporate a priori knowledge about *admissible data samples* into the learning process. These additional unlabeled data samples (called virtual examples or the *Universum*) are used along with labeled training samples, to perform an inductive inference. Note that the Universum samples are not real training samples; however they reflect a priori knowledge about application domain. For example, if the goal of learning is to discriminate between handwritten digits 5 and 8, one can introduce additional 'knowledge' in the form of other handwritten digits 0, 1, 2, 3, 4, 6, 7, 9. These examples from the Universum contain certain information about handwritten digits, but they cannot be assigned to any of the two classes (5 or 8). Detailed mathematical formulation of Universum learning, aka Universum SVM (U-SVM), can be found in [1,5,6].

Universum SVM formulation can be viewed as a generalization of standard SVM classification formulation. U-SVM setting is more complex, as it has more tunable parameters than standard SVM. So an important practical question is to formulate the conditions that enable improved prediction performance of U-SVM vs. standard SVM.

Recent papers [15] and [16] report such simple conditions for the effectiveness of Universum Learning for HDLSS setting. These conditions are based on analysis of the univariate histogram of projections. That is, the histogram of projections is generated first for standard SVM classifier, and then the shape of this histogram along with the distribution of projections of the Universum samples, is used to determine the effectiveness of the Universum for a given training set.

This technique is illustrated next using the same *MNIST handwritten digit data set*, where data samples represent handwritten digits 5 and 8, using RBF SVM. In this case, we

consider three different Universum data sets, i.e., handwritten digits 1, 3 and 6, and the problem is to evaluate relative effectiveness of these different types of Universa.

Cherkassky et al. [16] provide the following conditions for the effectiveness of a given Universum set for a particular labeled training set. That is, a Universum set is effective if its histogram of projections satisfies two conditions:

1. It has symmetric distribution relative to (standard) SVM decision boundary;
2. It has wide distribution between margin borders denoted as +1/-1 in the projected space.

Fig. 6 shows histograms of projections for 3 different types of Universa. Projections of labeled training data samples are shown in gray and black colors, and projections of Universum samples are shown in dashed-black. Histograms in Fig 6b and 6c seem to satisfy the conditions (1) and (2) better than in Fig. 6a. Thus digit 1 universum samples are not likely to provide significant improvement. Empirical comparison of the test error rate (shown in Table I) provided by standard SVM and U-SVM confirms this analysis.

TABLE I
TEST ERROR RATES FOR MNIST DATA WITH DIFFERENT UNIVERSA. TRAINING SET SIZE IS 1,000 SAMPLES. STANDARD DEVIATION OF REPORTED ERROR RATES IS GIVEN IN PARENTHESES

	SVM	U-SVM (digit 1)	U-SVM (digit 3)	U-SVM (digit 6)
Test error	1.47% (0.32%)	1.31% (0.31%)	1.01% (0.28%)	1.12% (0.27%)

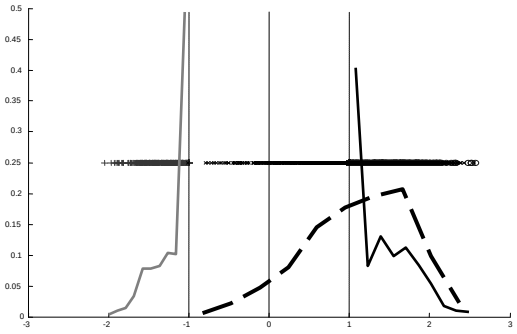


Fig 6a. Projections for digit 1 Universum

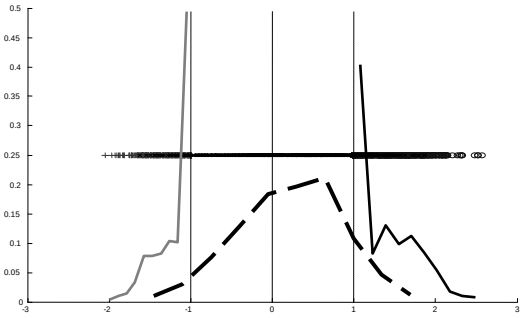


Fig. 6b. Projections for digit 3 Universum

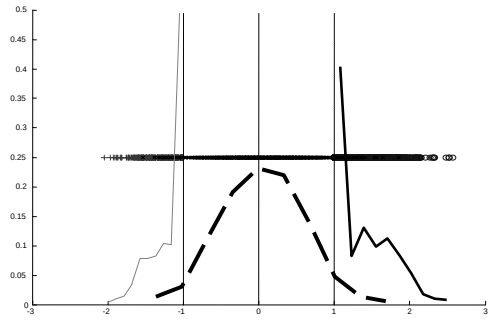


Fig. 6c. Projections for digit 6 Universum

Fig. 6 shows histograms of projections for 3 different types of Universa.

5 Summary

This paper presents a simple method for interpretation of high dimensional nonlinear SVM models in the form of univariate histogram of projections. This simple graphical technique can be used to understand the multivariate SVM models estimated from data under different standard and non-standard learning settings, including unbalanced data sets with unequal misclassification costs. This representation can also be used to explain the practical conditions for the effectiveness of new learning settings such as Universum SVM. Finally, such a simple graphical representation can be of immense help to practitioners and help them to have a better understanding of the SVM model.

6 References

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

[2] Hastie, T., R. Tibshirani and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, New York: Springer, 2001.

[3] Schölkopf, B. and A. Smola, *Learning with Kernels*. MIT Press, 2002.

[4] Camps-Valls, G., Rojo -Alvarez, J. L., and M. Martinez-Ramon, *Kernel Methods in Bioengineering, Signal and Image Processing*, London: Idea Group Publishing, 2007.

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

[6] Vapnik, V.N., *Statistical Learning Theory*. New York: Wiley, 1998.

[7] Caruana, R.,"Multi-task learning," *Machine Learning.*, vol. 28, pp. 41-75,July 1997.

[8] Evgeniou, T. and Pontil, M., "Regularized multi--task learning," in *Proc. 17th SIGKDD Conf. on Knowledge Discovery and Data Mining*,2004, pp. 109-117.

[9] Liang, L. and Cherkassky, V.,"Connection between SVM+ and Multi-Task Learning," *IJCNN*, 2008.

[10] Doina Caragea, Dianne Cook, Vasant G. Honavar, "Gaining insights into support vector machine pattern classifiers using projection-based tour methods," SIGKDD 2001, 251-256.
[11] Thanh-Nghi Do, François Poulet, "Enhancing SVM with Visualization," Discovery Science 2004, 183-194.
[12] A. Jakulin, M. Možina, J. Demšar, I. Bratko and B. Zupan,"Nomograms for visualizing support vector machines," Conference on Knowledge Discovery in Data 2005, 108—117.
[13] Sam Roweis, "sam roweis:data,"[Online].Available: http://www.cs.nyu.edu/~roweis/data.html[Accessed: May 5, 2010] .
[14] Feng, C., Cherkassky, V., Weisdorf, D., Arora, M., Van Ness, B., "Predictive modeling of Transplant-Related Mortality," *Proc. of the 2010 Design of Medical Devices Conf.*, Minneapolis, April 2010.
[15] Cherkassky, V. and W. Dai, "Empirical Study of the Universum SVM Learning for High-Dimensional Data," *in Proc. ICANN*, 2009.
[16] Cherkassky,V. , S. Dhar and W. Dai, "Practical Conditions for Effectiveness of the Universum Learning," IEEE Trans. on Neural Networks, Feb 2010, submitted.