

Gender Classification of Human Faces Using Inference through Contradictions

Xue Bai and Vladimir Cherkassky

Abstract—We present an empirical study of gender classification of human faces, using new learning methodology called inference through contradictions, introduced in [9]. This approach allows to incorporate a priori knowledge in the form of additional (unlabeled) samples, called the Universum, into the supervised learning process. Application of this methodology to gender classification shows that using this approach enables better generalization over standard SVM classification (using labeled data alone).

I. INTRODUCTION

THERE is a growing need for development of powerful and robust methods for estimating models from data. In many applications, good models are defined in terms of their generalization capability, where the goal is to estimate unknown dependency from historical (training) data $(\mathbf{x}_i, y_i), (i = 1, \dots, n)$, in order to use this model for predicting future (test) samples. Most supervised learning methods developed in statistical learning, pattern recognition and machine learning are based on standard inductive formulation of the learning problem [8] [4] [6].

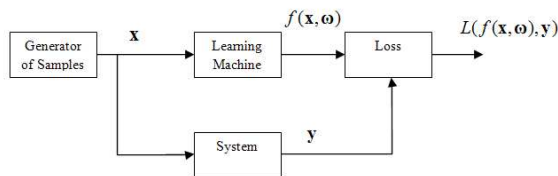


Fig. 1. Generic system for inductive learning.

Under inductive formulation (see Fig. 1), the learning machine observes a finite number of i.i.d. samples, aka training data, and the goal of learning is to estimate unknown mapping $f : \mathbf{x} \rightarrow y$ in order to imitate the system's response for future inputs (aka test data). This function is selected from a set of admissible functions (or approximating functions) $f(\mathbf{x}, \omega)$ implementable by the Learning Machine. For regression problems, the system's output is real-valued $y \in \mathcal{R}$ and for (binary) classification problems, the output is class label $y \in \{+1, -1\}$. Hence, the problem of learning can be stated as function estimation using finite (training) data. The quality of 'useful' models is measured in terms of their

generalization capability, i.e. minimization of the prediction risk functional (measuring 'generalization'). The learning system shown in Fig. 1 suggests that the goal of learning is to 'imitate' the output of unknown system, as expressed in the goal of minimization of a given loss function. This goal of *system imitation* is different from the goal of *system identification* (or density estimation) adopted in classical statistics and function approximation theory ([10], [5]).

The usual assumptions (behind inductive learning) may not hold for many applications. For example, if the input values of the test samples are known (given), then an appropriate goal of learning may be to predict outputs *only* at these points. This leads to transduction formulation [8].

Vapnik-Chervonenkis theory makes a strong argument that for finite sample estimation problems one should always use the most appropriate *direct formulation* of the learning problem [8]. This principle can be also applied on the level of formalizing application-domain requirements ([5],[3],[2]). That is, for a given application, one should first introduce an appropriate learning problem formulation (reflecting application domain requirements), and only then develop learning algorithms appropriate for this learning setting.

The practical need for non-inductive formulations can be further motivated by sparse high-dimensional data common in many applications, i.e.

- genetic micro-array data analysis, where many gene expression levels have been measured for a few cases.
- medical imaging, where a small number of 2D or 3D images are represented by vectors of many parameters. For example, functional magnetic resonance imaging (fMRI) is concerned with analysing a few hundred of 3D images (training sample size $n \sim 100$) of very high dimensionality (the number of voxels $d \sim 10,000$).
- text or document categorization, where documents are represented as a high-dimensional feature vectors, so that the presence (or absence) of a particular word in a document is encoded as 1 (or 0) entry in the feature space ($d \sim 10,000$).

Most approaches to learning with high-dimensional data focus on improvements to *existing inductive methods* (i.e., LDA or SVM) that try to incorporate a priori knowledge about the good models (i.e., via specially designed SVM kernels). These approaches, however, are fundamentally constrained by the inductive learning setting itself. In contrast, *non-inductive learning* methodologies focus on the most appropriate *direct formulation* of the learning problem. It can be argued that most recent advances in statistical learning (i.e., transduction, semi-supervised learning, single-class learning,

Xue Bai and Vladimir Cherkassky, IEEE Fellow are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55414 (email: {baixx015,cherk001}@umn.edu).

This work was supported, in part, by NSF grant ECCS-0802056.

multi-task learning) reflect an improved understanding of the learning problem setting. In this paper we investigate application of a new learning setting, called *Learning through Contradictions* [9], that allows to improve prediction accuracy of a classifier, by incorporating a priori knowledge in the form of additional data samples, into the learning process. For example, for the well-known problem of *handwritten digit recognition*, one can use additional a priori knowledge in the form of examples of handwritten letters. These handwritten letters reflect the style of writing and thus can potentially improve generalization of the classifier (for test digits). This leads to the setting *Learning through Contradictions*, or learning in the Universum environment ([9], [10]). Recent study [11] shows that Learning through Contradictions can improve classification performance for handwritten digit recognition and several other applications.

This paper describes an application of this new learning methodology to gender classification of human faces. The paper is organized as follows. Section 2 describes mathematical formulation of Learning through Contradictions ([10]). Section 3 presents empirical results for gender classification. Summary is given in section 4.

II. INFERENCE THROUGH CONTRADICTIONS

The idea of 'inference through contradictions' was introduced by Vapnik [9] in an attempt to introduce a priori knowledge into the learning process. Recall that all traditional approaches (for encoding a priori knowledge) try to characterize the *space of admissible models* $f(x, \omega)$, or the relationship between the 'true' model and the properties of admissible models. This includes, for example, specification of kernels (in SVM), or specification of prior distributions (in Bayesian methods). It may be argued that in real applications (especially with sparse high-dimensional data) such 'good' parameterizations are hard to come by. So it is more reasonable to introduce a priori knowledge about *admissible data samples*. These additional unlabeled data samples (called virtual examples or the *Universum*) are used along with labeled training samples, to perform an inductive inference. Examples from the Universum are not real training samples, however they reflect a priori knowledge about application domain. For example, in the problem of hand-written digit recognition, one can introduce virtual examples in the form of handwritten letters. These examples from the Universum reflect 'style of writing' but they can not be assigned to any of the classes (digits). Effectively, the Universum data contains a priori knowledge about the *region of the input space* where the data is likely to belong.

The idea of using additional data to improve learning is not new. However, the Universum data is different from additional data used in earlier methods. Additional labeled data (also called 'virtual examples') is used in the method of 'hints' [1]; however, these 'hints' are used to encode a priori knowledge about the properties of good models, rather than knowledge about the input space. Likewise, in SVM knowledge about the properties of good models is used in the Virtual Support Vectors method [7].

Let us consider the inductive setting (for binary classification), where we have labeled training data (x_i, y_i) , $(i = 1, \dots, n)$, and a set of unlabeled examples from the Universum, (x_j^*) , $(j = 1, \dots, m)$. The Universum contains data that belongs to the same application domain as the training data, but these samples are *known not to belong* to either class. The main question is: how to incorporate the Universum samples into inductive learning? Next we explain how it can be done using informal arguments, for binary classification problem. Let us assume that labeled training data is linearly separable using large margin hyperplanes $f(x, \omega) = (x \cdot \omega) + b$ (as in standard SVM). Then the Universum samples can either fall *inside* the margin or *outside* the margin (see Fig. 2). Recall that the Universum samples do not belong to either class, so we favor hyperplanes with Universum samples inside the margin. Such Universum samples (inside the margin) are called *contradictions*, because they are falsified by the model (i.e., have non-zero slack variables for either class label). So the new mode of inference implements a trade-off between explaining training samples (using large-margin hyperplanes) and maximizing the number of contradictions (on the Universum). More formally, this new mode of inference implements Structural Risk Minimization (SRM) structure where each element (of a structure) is indexed by the number of contradictions ([10]). This can be contrasted to standard SVM classification, which implements a structure indexed by the size of margin.

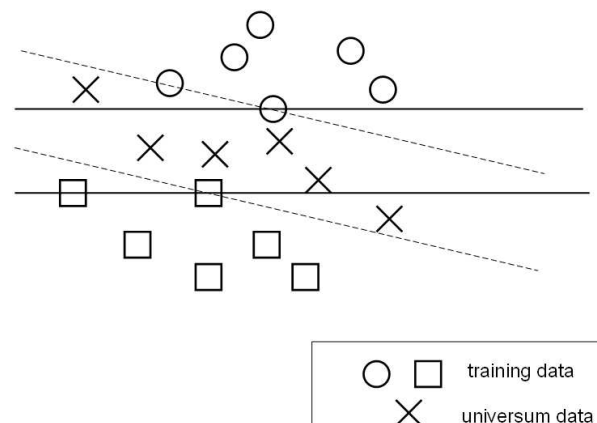


Fig. 2. Two large-margin separating hyperplanes explain training data equally well, but have different number of contradictions on the Universum. The model with a larger number of contradictions is favored.

The quadratic optimization formulation for implementing SVM-style inference through contradictions is shown next following ([10]). For labeled training data, we use standard SVM soft-margin loss with slack variables ξ_i . For the Universum samples (x_j^*) , we need to penalize the real-valued outputs of our classifier that are 'large' (far away from zero). So we adopt ϵ -insensitive loss (as in standard SV regression). Let ξ_j^* denote slack variables for samples from the Universum. Then SVM-based inference through

contradiction can be stated as follows:

$$\text{minimize } R(\omega, b) = \frac{1}{2}(\omega \cdot \omega) + C \sum_{i=1}^n \xi_i + C^* \sum_{j=1}^m \xi_j^*,$$

where $C, C^* \geq 0$, subject to constraints
 $y_i[(\omega \cdot \mathbf{x}_i) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, n$
(for labeled data)
 $|(\omega \cdot \mathbf{x}_i) + b| \leq \epsilon + \xi_j^*, \quad \xi_j^* \geq 0 \quad j = 1, \dots, m$
(for the Universum), where $\epsilon \geq 0$

(1)

Parameters C and C^* control the trade-off between minimization of errors and maximizing the number of contradictions. Selecting 'good' values for these parameters is a part of model selection (usually performed via resampling). When $C^* = 0$, this formulation is reduced to standard soft-margin SVM. Solution to the above optimization problem defines the large margin hyperplane $f(\mathbf{x}, \omega^*) = (\mathbf{x} \cdot \omega^*) + b^*$ that incorporates a priori knowledge (data from the Universum) into the final SVM model. The dual formulation for inductive SVM in the Universum environment, and its nonlinear kernelized version can be readily obtained using standard optimization theory and SVM techniques [10]. The above quadratic optimization problem is convex due to convexity of constraints for labeled data and for for the Universum. Efficient computational algorithms for solving this optimization problem involve minor modifications of standard SVM software [11].

III. GENDER CLASSIFICATION OF HUMAN FACES

This section describes a challenging application, gender classification, illustrating advantages of learning in the Universum environment with sparse data. The application itself involves learning of gender (male or female) from face images. This is a difficult problem, due to high variability of face data (i.e., eyeglasses, hairstyle, facial expression, etc), and sparseness of data. That is, the number of labeled training examples is very small (13 training samples used in this study). The purpose of this study was to show relative advantages of using this new learning methodology (in comparison with standard SVM classifier), and also to investigate several possibilities for generating Universum data. The main issue for implementing inference through contradictions is specifying 'good' Universum data. Even though this process can not be formalized, existing empirical studies suggest that it should be easy to collect or generate Universum data for most applications. Several methods for generating the Universum data have been proposed for handwritten digit recognition [11], but it is not clear whether they also apply to the problem of gender classification. This study used publicly available universum SVM software obtained from <http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html>

A. Data Set and Preprocessing

Face image data was obtained from *Face Recognition Data, University of Essex, UK*, available at <http://cswww.essex.ac.uk/mv/allfaces/index.html>. It contains 24-bit color headshots of 32 males and 20 females, 5 images

per person. See Fig. 3. Note that all face images are centered and are (approximately) in the same scale. Preprocessing involved standard histogram equalization for each image. All images are converted to 256-level gray-scale and to the same size of 45×50 pixels. Therefore, each image can be represented as a 2250 dimensional real-valued vector.

The data base contains total of 260 images of human faces of 52 individuals (5 photos per person). There are total of 20 males and 32 females, labeled -1 and $+1$ respectively.



Fig. 3. Generic system for inductive learning.

B. Experimental Set-Up

A set of 52 individuals (20 males and 32 females) is randomly split into 4 equal-size partitions, by selecting 5 females and 8 males for training, and the remaining individuals for testing. For each individual, a single image is then randomly selected (out of 5 photos). So the size of the training set is 13, and the size of the test set is 39. The classifier is then trained using training data, and its prediction accuracy is estimated using test data. For every training/test data partition, the classification accuracy obtained by a standard SVM classifier is compared to that of the Universum SVM classifier.

To reduce variability due to random data, for each partitioning of 52 individuals (as specified above), the experiment was repeated 10 times, by randomly selecting a different photo for each person. The average classification accuracy and its standard deviation, for each partitioning, are then reported.

C. SVM Parameter Tuning

Linear SVM classifier was used in this study, due to high-dimensionality and sparseness of the training data. Various combinations of (C, C^*) are tested ($C = 10^{-1}, 1, 10, C^* = 10^{-3}, 10^{-2}, 10^{-1}, 1$). Empirically we found that good parameter values $C = 1$ (for linear SVM classifier), and $C = 1, C^* = 1$ (for the Universum SVM) are very insensitive to variations in the input data, and to the size of the Universum data set. So these (fixed) values have been used in all experimental comparisons.

D. Generation of the Universum Data

Three different possibilities for generating Universum data were investigated:

- 1) *Random Average*. Randomly sample one male and one female training sample, and compute the average. See Fig. 4.
- 2) *Empirical Distribution*. Following the method by [11], we use the intensity values of every pixel (x, y) of the training data, in order to estimate the two-dimensional distribution of pixel intensities. Then Universum samples are generated from this empirical distribution.
- 3) *Animal faces* (Fig. 5). We collected 50 pictures of animal 'faces' and manually cropped each of them so that the face approximately covers the same amount of portion of the image as the human faces. The pictures were also converted and resized in the same format as the human face images. For Universum set sizes larger than 50, artificial samples were generated by computing pixel-wise averages of randomly chosen animal faces.

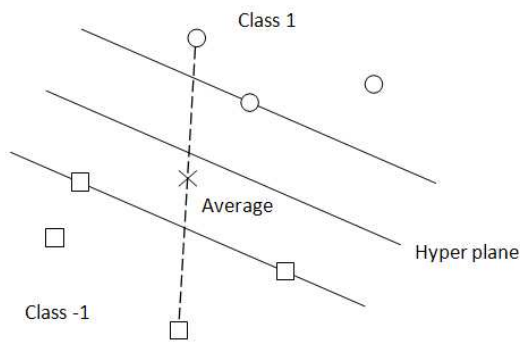


Fig. 4. Generation of the Universum data by averaging.



Fig. 5. Examples of animal faces.

E. Comparison Results

Experimental results are summarized below in the form of Table I and pie charts (Fig. 6 and 7) showing classification accuracy (on an independent test set) obtained using standard linear SVM vs the Universum SVM with a different number of Universum samples ($N = 100, 500, 1000$). Comparisons are presented separately for each of the 4 partitioning of the available data into training/test subsets (as described under *Experimental Set-Up* sub-section). These results indicate the

2 – 3% improvement in the classification accuracy (over standard inductive SVM) with the Universum data generated via random averaging, and very minor improvement ($\sim 1\%$) with the Universum data generated via empirical distribution. Using 'animal faces' as the Universum data actually decreases the classification accuracy by two to five percent (vs standard SVM. See Table II), suggesting that animal faces are not relevant to the problem of interest (human faces).

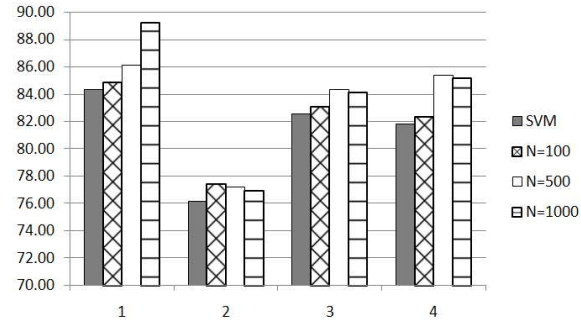


Fig. 6. Universum generation: random averaging.

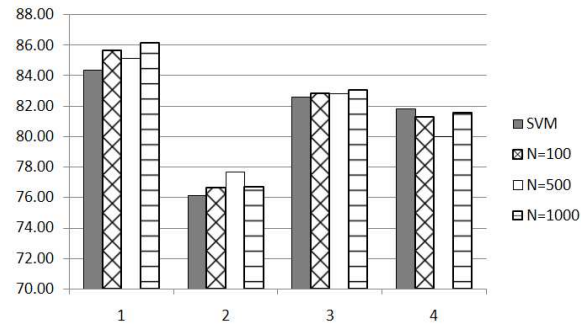


Fig. 7. Universum generation: empirical distribution.

TABLE I
SUMMARY OF COMPARISON RESULTS. AVERAGE(STANDARD DEVIATION)

Partition		P1	P2	P3	P4
SVM		84.4 (5.5)	76.2 (6.0)	82.6 (3.6)	81.8 (4.1)
Avg.	100	84.9 (5.7)	77.4 (4.5)	83.1 (3.5)	82.3 (4.4)
	500	86.2 (3.5)	77.2 (5.7)	84.4 (3.3)	85.4 (3.6)
	1000	89.2 (2.4)	76.9 (5.1)	84.1 (3.4)	85.1 (2.9)
Emp.	100	85.6 (2.8)	76.7 (4.8)	82.8 (3.0)	81.3 (3.6)
	500	85.1 (2.9)	77.7 (4.2)	82.8 (3.8)	81.3 (3.8)
	1000	86.2 (5.7)	76.7 (4.6)	83.1 (4.4)	81.5 (2.9)

IV. SUMMARY

Our empirical results show that using learning through contradictions with Universum examples obtained via random averaging can effectively improve classification performance. The relative improvement is best when the number

TABLE II
COMPARISON RESULTS: USING ANIMAL FACES AS UNIVERSUM

Partition	P1	P2	P3	P4
SVM	82	80	87	84
50	78	71	81	78
100	80	71	82	77
500	81	70	82	75

of Universum samples is large (500 or 1000). This finding is consistent with previous study [11]. Note that an improvement was observed for all 4 partitionings of the data (see Fig. 6). As noted by [10], improved generalization can be explained by the fact that the Universum formulation effectively enables solving the SVM problem in a lower-dimensional manifold defined by the Universum data. In other words, appropriately chosen Universum data implicitly defines the subspace of the input space where the data live. Proper selection of 'good' Universum data is application-dependent and can not be formalized. Notably, even for non-optimal selection of the Universum (i.e. via empirical distribution), the generalization performance is quite robust and stable (see Fig. 7). Generation of the Universum via empirical distribution of pixel intensities is non-optimal because individual pixels are assumed to be independent (of each other).

However, when the Universum data is chosen inappropriately (i.e., animal faces), the generalization performance actually degrades. Overall, our findings are consistent with an earlier study [11], in that:

- learning with the Universum can improve generalization performance, especially when the number of training samples is small. (recall that our study uses just 13 labeled samples);
- poorly chosen Universum (animal faces) actually degrades classification performance (Table II).

REFERENCES

- [1] Y. S. Abu-Mostafa. Hints. *Neural Computation*, 7(4):639–671, 1995.
- [2] V. Cherkassky. New formulations for predictive learning. *Invited Lecture, ANNIE-2004*, St. Louis MO, 2004.
- [3] V. Cherkassky. Alternative formulations for predictive learning. *Plenary Lecture, International Conference on Artificial Neural Networks (ICANN)*, Vienna, Austria, 2001.
- [4] V. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods*. Wiley, New York, 1998.
- [5] V. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods, Second Edition*. Wiley, New York, 2007.
- [6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
- [7] B. Scholkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.
- [8] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
- [9] V. N. Vapnik. *Statistical learning theory*. Wiley, New York, 1998.
- [10] V. N. Vapnik. *Estimation of Dependencies Based on Empirical Data. Empirical Inference Science: Afterword of 2006, Second Edition*. Springer, 2006.
- [11] J. Weston, R. Collobert, F. Sinz, L. Bottou, and V. Vapnik. Inference with the universum. *Proceedings of ICML*, Pittsburgh, USA, 2006.