INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET)

ISSN 0976 – 6367(Print)
ISSN 0976 – 6375(Online)
Volume 5, Issue 12, December (2014), pp. 226-233
© IAEME: www.iaeme.com/IJCET.asp
Journal Impact Factor (2014): 8.5328 (Calculated by GISI)
www.jifactor.com

IJCET

©IAEME

UNIFIED APPROACH OF GENERATIVE AND DISCRIMINATIVE LEARNING

Varsha S. Futane

[1]PG Student (ME-CSE) Computer Sciences and Engineering Department
TPCT's College of Engineering, Osmanabad, (M.S.), India

Dr. Anilkumar N. Holambe

[2]Professor and Head Computer Sciences and Engineering Department
TPCT's College of Engineering, Osmanabad, (M.S.), India

ABSTRACT

Generative and discriminative learning are two of the major paradigms for solving prediction problems in machine learning, each offering important distinct advantages. They have often been studied in different sub-communities, but over the past decade, there has been increasing interest in trying to understand and leverage the advantages of both approaches. The goal of this paper is to map out our current understanding of the empirical and theoretical advantages of each approach as well as their combination, and to identify open research directions.

Keywords: Algorithms, Generative Learning, Discriminative Learning, Machine Learning Semi-Supervised, Unlabelled Data, Machine Learning

I. INTRODUCTION

Machine learning techniques are now widely used in computer vision. In many applications the goal is to take a vector **x** of input features and to assign it to one of a number of alternative classes labelled by a vector **c** (for instance, if we have *C* classes, then **c** might be a *C*-dimensional binary vector in which all elements are zero except the one corresponding to the class). Throughout this paper we will have in mind the problem of object recognition, in which **x** corresponds to an image (or a region within an image) and **c** represents the categories of object (or objects) present in the image, although the techniques and conclusions presented are much more widely applicable. In the simplest scenario, we are given a training data set comprising *N* images **X** = {**x**1, . . . , **x**N} together with corresponding labels **C** = {**c**1, . . . , **c**N}, in which we assume that the images, and their labels, are drawn independently from the same fixed distribution. Our goal is to predict the class $\hat{\mathrm{C}}$ for a new input vector $\hat{\mathrm{X}}$, and so we require the conditional distribution

$$p(\widehat{\mathrm{C}}|\widehat{\mathrm{X}}, \mathrm{X}, \mathrm{C}) \qquad (1)$$

To determine this distribution we introduce a parametric model governed by a set of parameters $\boldsymbol{\theta}$. In a discriminative approach we define the conditional distribution $p(\mathbf{c}|\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of the model. The likelihood function is then given by

$$L\,(\theta) = p(C|X,\theta) = \prod_{n=1}^{N} p(c_n|x_n,\theta) \qquad (2)$$

If training data is plentiful a point estimate for $\boldsymbol{\theta}$ can be made by maximizing the posterior distribution to give $\boldsymbol{\theta}_{\mathrm{MAP}}$, and the predictive distribution then estimated using

$$p(\hat{c}|\hat{x}\;X,C) = p(\hat{c}|\hat{x}\;\theta_{MAP}) \qquad (3)$$

In recent years there has been growing interest in a complementary approach based on generative models, which define a joint distribution $p(\mathbf{x}, \mathbf{c}|\boldsymbol{\theta})$ over both input vectors and class labels [1].

In this paper we develop a unified approach which says that, for a given model, there is a unique likelihood function and hence there is only one correct way to train it. The 'discriminative training' of a generative model is instead interpreted in terms of standard training of a different model, corresponding to a different choice of distribution. This removes the apparently ad-hoc choice for the training criterion, so that all models are trained according to the principles of statistical inference. Furthermore, by introducing a constraint between the parameters of this model, through the choice of prior, the original generative model can be recovered.

II. A NEW VIEW OF DISCRMINATIVE LEARNING

A parametric generative model is defined by specifying the joint distribution $p(\mathbf{x}, \mathbf{c}|\boldsymbol{\theta})$ of the input vector $\mathbf{x}$ and the class label $\mathbf{c}$, conditioned on a set of parameters $\boldsymbol{\theta}$. Since the data points are assumed to be independent, the joint distribution is given by

$$L\mathrm{G}(\boldsymbol{\theta}) = p(\mathbf{X},\mathbf{C}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(x_n c_n|\theta) \qquad (4)$$

In order to improve the predictive performance of generative models it has been proposed to use 'discriminative training' [2] which involves maximizing

$$L\mathrm{D}(\boldsymbol{\theta}) = p(\mathbf{C}, \boldsymbol{\theta}|X) = p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(c_n|x_n,\theta) \qquad (5)$$

in which we are conditioning on the input vectors instead of modelling their distribution. Here we have used

$$p(c|x.\theta) = \frac{p(x,c|\theta)}{\sum_{c'} p(x,c'|\theta)} \qquad (6)$$

Note that (5) is not the joint distribution for the original model defined by (2), and so does not correspond to MAP for this model. The terminology of ‘discriminative training’ is therefore misleading, since for a given model there is only one correct way to train it. It is not the training method which has changed, but the model itself.

Following [3] we therefore propose an alternative view of discriminative training, which will lead to an elegant framework for blending generative and discriminative approaches. Consider a model which contains an additional independent set of parameters $\tilde{\theta} = \{\breve{\Pi}, \tilde{\lambda}\}$ in addition to the parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\lambda}\}$, in which the likelihood function is given by

$$q(\mathbf{x}, \mathbf{c}|\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = p(\mathbf{c}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\widetilde{\boldsymbol{\theta}}) \qquad (7)$$

Now suppose instead that we consider a prior which enforces equality between the two sets of parameters.

$$p(\boldsymbol{\theta},\tilde{\theta}) = p(\boldsymbol{\theta})\delta(\boldsymbol{\theta} - \tilde{\theta}) \qquad (8)$$

Then we can set $\tilde{\theta} = \boldsymbol{\theta}$ in (7) from which we recover the original generative model $p(\mathbf{x}, \mathbf{c}|\boldsymbol{\theta})$. Thus we have a single class of distributions in which the discriminative model corresponds to independence in the prior, and the generative model corresponds to an equality constraint in the prior.

III. BLENDING GENERATIVE AND DISCRIMINATIVE

Clearly we can now blend between the generative and discriminative extremes by considering priors which impose a soft constraint between $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. The benefit of the generative approach is that it can make use of unlabelled data to augment the labeled training set. Suppose we have a data set comprising a set of inputs **X**L for which we have corresponding labels **C**L, together with a set of inputs **X**U for which we have no labels. For the correctly trained generative model, the function which is maximized is given by

$$[p(\theta) \prod_{n\epsilon L} p(x_n, c_n|, \theta) \prod_{m\epsilon U} p(x_m|, \theta)] \qquad (9)$$

We see that the unlabelled data influences the choice of $\boldsymbol{\theta}$ and hence affects the predictions of the model. By contrast, for the ‘discriminatively trained’ generative model the function which is now optimized is again the product of the prior and the likelihood function and so takes the form

$$p(\boldsymbol{\theta}) = \prod_{n\epsilon L}^{N} p(x_c|x_n, \theta) \qquad (10)$$

and we see that the unlabelled data plays no role. Thus, in order to make use of unlabelled data we cannot use a discriminative approach. Now let us consider how a combination of labelled and unlabelled data can be exploited from the perspective of our new approach for which the joint distribution becomes

$$q(\mathbf{X}\mathrm{L},\mathbf{C}\mathrm{L},\mathbf{X}\mathrm{U}, \boldsymbol{\theta}, \tilde{\theta}) = p(\boldsymbol{\theta},\tilde{\theta})\left[\prod_{n\epsilon L}^{N} p(c_n|x_n, \theta)p(x_n|\tilde{\theta})\right]\left[\prod_{m\epsilon U}^{N} p(x_m|\tilde{\theta})\right] \qquad (11)$$

We see that the unlabelled data (as well as the labelled data) influences the parameters $\tilde{\theta}$which in turn influence $\boldsymbol{\theta}$ via the soft constraint imposed by the prior.

IV. EXPERIMENTS AND RESULTS

4.1 Feature Extraction

We will follow several recent approaches and use interest point detectors to focus attention on a small number of local patches in each image. This is followed by invariant feature extraction from a neighbourhood around each interest point. We choose to work with Harris-Laplace (HL) [4] and Difference of Gaussian (DoG) [5] interest point operators because they are invariant to orientation and scale changes. In earlier study [6] we have used DoG interest point detector with SIFT (Scale Invariant Feature Transform) descriptor. SIFT is invariant to illumination and affine (to some degree) changes and very suitable for DoG interest point detectors. However SIFT, being a 128 dimensional vector, brings a high computational load for model learning. Thus, in this section we will use 15 dimensional Local Jet (LJ) descriptor instead [7,8]. For the purpose of comparison, we will train our models using different feature types and see how they are affected by these choices. The two feature point operators, HL and DoG, will be used with the same feature descriptor (LJ).

The feature descriptor may be concatenated with colour information. The colour information is extracted from each patch based on [1]. Averages and standard deviations of (R,G,B), (L, a, b) and ($r = R/(R+G+B)$, $g = G/(R+G+B)$) constitute the colour part of the feature vector. Lab is a device-independent colour space that attempts to uniformly represent colour as we perceive it. L is the lightness value, a is the red/green opponency and blue/yellow is represented on the b axis. As a result, if colour is also used as a feature descriptor then we will have a 31 dimensional feature vector. The number of random patches is selected to be approximately the same as the number of patches found by other interest point operators, which is around 100 for each image. In the rightmost image in Figure 1 the cow image with some of the random patches is also shown.. In Section 4.2, comparison of the two models when used with different features will be given in terms of patch labelling and image labelling. We will compare HL and DoG operators with LJ and colour feature, and random patches with PCA coefficients and colour feature.

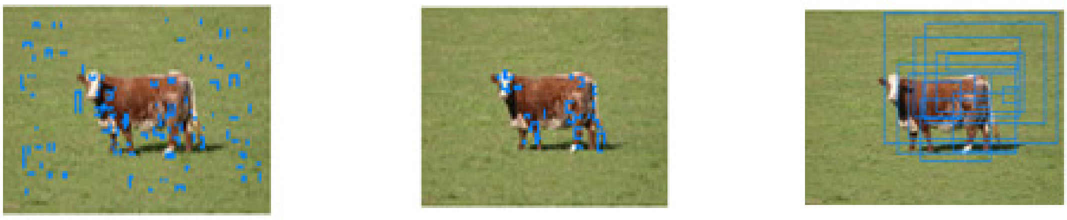


Fig. 1: Different interest point operators. Feature point locations are the centers of the squares and the size of a square shows the scale of that feature point. The three images show (left to right) DoG interest points, HL interest points and random patches.

In this section, we have used a test bed of weakly labelled images each containing cows, in which the animals vary widely in terms of number, pose, size, colour and texture. There are 167 images in each class, and 10-fold cross-validation is used to measure performance. For the generative model we used a separate Gaussian mixture for cow, and background, each of which has 10 components with diagonal covariance matrices in our earlier study [6] we used input vector of size 144 which consists of sift and colour features using a smaller feature vector this time brings computational benefit such as speed and computable covariance matrixes. In Figure 2 examples for generative model patch labelling are given for different situations where most probable label is assigned for each patch. Patch centers are shown by coloured dots where colour denotes the class (red, white, green for cow and background respectively).

4.2 Comparison with Different Feature Types

In this section we will provide comparative results between our generative (G)and soft discriminative (D) model when they are used with different types of features such as HL operator with LJ and colour feature (HL-LJ+C), DoG operator with LJ and colour (DoG-LJ+C) and random patches with PCA coefficients and colour feature (R-PCA+C). Usually DoG feature point operator finds more points than HL operator does when applied on the same image. In the random selection case we define the number of feature points and their local extension.

In order to eliminate the effect of data quantity in the comparison, we arranged the feature point extraction algorithms so that they produce roughly the same amount of feature points (around 100) for each image. Means and standard deviations of overall correct rate results over 10 fold runs are given in Table 1.

Table 1. Means (M) and standard deviations (SD) of overall correct image label rate for different feature types: HL with LJ and colour (HL-LJ+C), DoG with LJ and colour (DoG-LJ+C) and random patches with PCA coefficients and colour (R-PCA+C)

	HL-LJ+C	DoG-LJ+C	R-PCA+C
D (M)(%)	80.63	89.38	78.13
D (SD)(%)	7.13	4.74	3.83
G (M)(%)	56.25	56.25	75.62
G (SD)(%)	6.25	9.88	2.61

Columns are for different feature types and rows are for different models. The best overall correct rate for the discriminative model is obtained by DoG-LJ+C feature and R-PCA+C feature causes the worst performance. The generative model produces highly different overall correct rates with different feature types. The best performance for the generative model is obtained by the random patches. With DoG-LJ+C and HLLJ+C the performance is worse than the random patches.

It is also interesting to investigate the extent to which the discriminative and generative models correctly label the individual patches. In order to make a comparison in terms of patch labelling we use 6 hand segmented test images for each class. These segmented images are different from those we have used for initializing and training the models. Normalization is required for the discriminative model in order to obtain patch label probabilities.

Various thresholds are used on patch label probabilities in order to produce ROC curves for the generative model and the soft discriminative model, as shown in Figure 3.As can be observed from the plots the generative model patch labelling is better than the discriminative model patch labelling for all types of features and patch labelling with DoG operator with LJ and colour feature is better than other feature types.

Some examples of patch labelling for test images are given in Figure 4 for random patches and for DoG patches, and in Figure 5 for HL patches. In these figures each patch is assigned to the most probable class and patch centers are given with coloured dots where colour denotes the patch label.

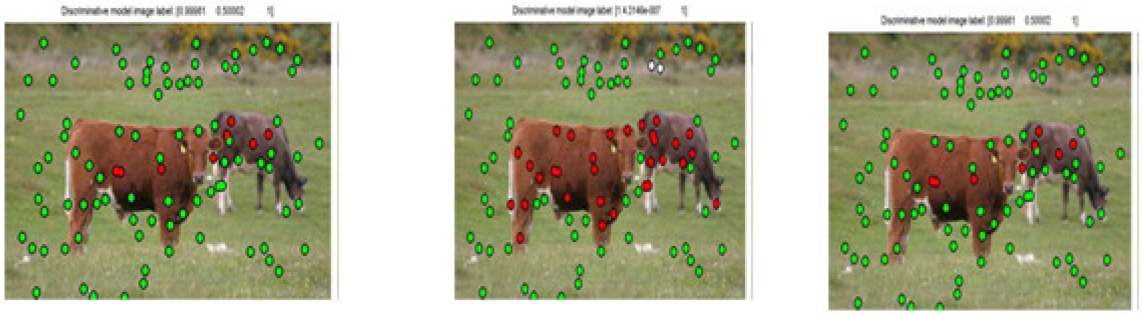


Fig. 2: Patch labelling results (red, white, green for cow and background respectively). Leftmost image is obtained when segmented data is not used in training of probabilistic noisy OR discriminative model. Middle image is when segmented data is used for training the same model. The rightmost image is when the soft discriminative model is trained with segmented data.

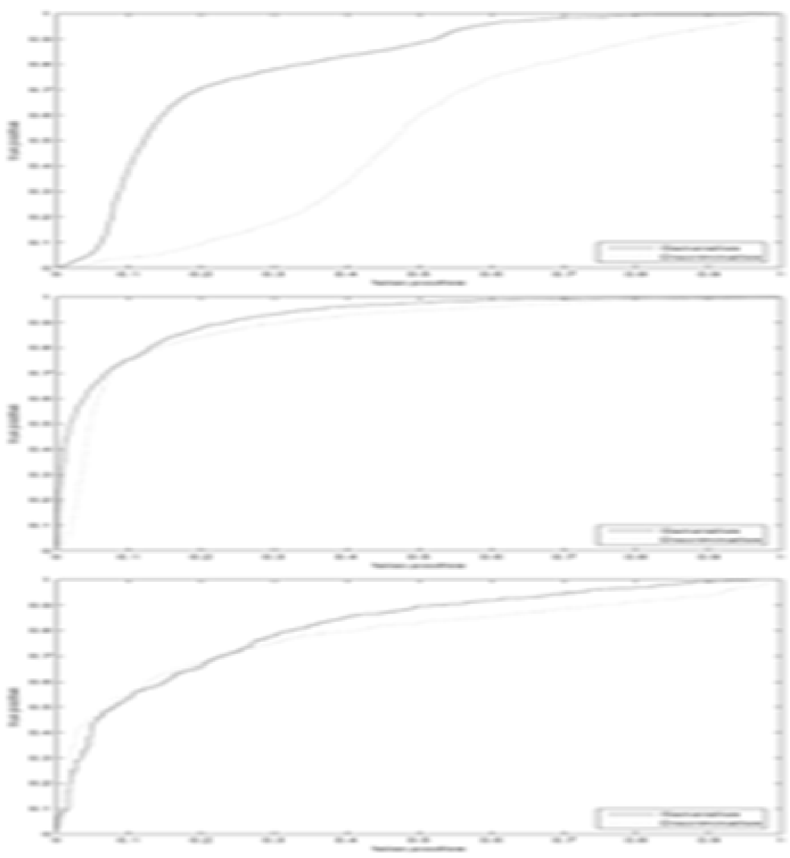


Fig. 3. ROC curves of patch labelling. Each figure contains two curves. One for the generative model and the other one for the discriminative model. Upper figure is for R-PCA+C patches. Center one is for DoG-LJ+C. Bottom one is for HL-LJ+C.

4.3 Comparison for Training Data Quantity

We trained our models with various number of training data. Similar results as [10] and [9] are obtained in this section also. Since the generative model performs the best with random patches we were expecting that with less data the generative model performance should be better than discriminative model. As can be observed from the left plots in Figure 6 the generative model performance is much better than the discriminative one for less data and as the quantity of data is increased discriminative model performance increases much faster than the generative model's performance. When DoGLJ+ C features are used, since the generative model does not perform well with this feature type, we were not expecting same type of behaviour. As can be seen in the right hand plots in Figure 6, the generative and the discriminative models behave nearly the same as we increase the data quantity but the discriminative model performs better than the generative model all the time.

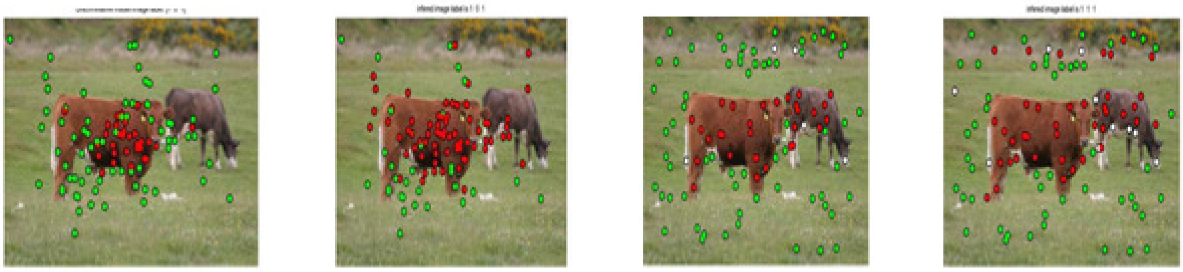
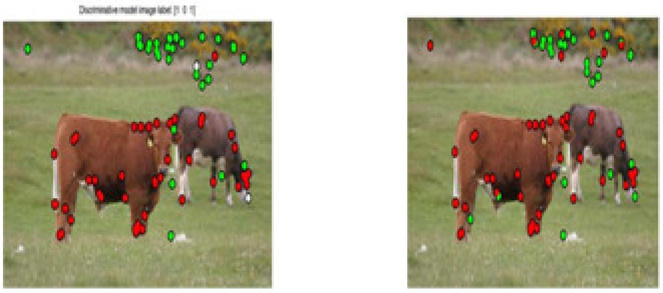


Fig. 4. Patch labelling examples for random patches (a) and for DoG patches (b)



(a) **(b)**

Fig.5. Patch labelling examples for HL patches. Results for discriminative model (a) and generative model (b)

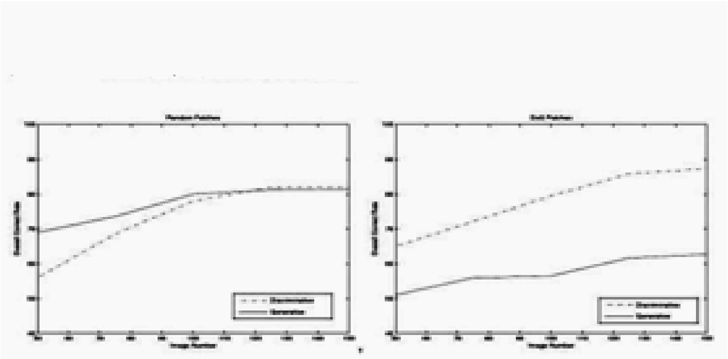


Fig. 6. Overall correct rate versus data number plots to show how the models behave as the data quantity is increased. Left figure is when random patches are used and the right figure is when DoG features are used.

4.4 Discussion

We have compared the two models when different number of images are used for training. When this comparative experiment is performed using random patches as features, we have observed that with small number of data the generative model performs better than the discriminative model and as the data quantity increases the performances for both models increase but this increase is more marked for the discriminative model, so that the performance of the two approaches is similar for large data sets. When this comparative experiment is performed using DoG-LJ+C features, both models behaved nearly the same for all data quantities but the discriminative model performs better all the time as we increase the data quantity.

V. CONCLUSION

In this paper we have shown that 'discriminative training' for generative models can be re-cast in terms of standard training methods applied to a modified model. Although we have focused on a specific application in computer vision concerned with object recognition, the techniques proposed here have very wide applicability.

A unified approach to combining generative and discriminative learning not only gives a more satisfying foundation for the development of new models, but it also brings practical benefits. This new viewpoint opens the door to a wide range of new models which interpolate smoothly between generative and discriminative approaches and which can benefit from the advantages of both.

REFERENCES

[1] T. Jebara. Machine Learning: Discriminative and Generative. Kluwer, 2004.

[2] O.Yakhnenko, A. Silvescu, and V. Honavar. Discriminatively trained markov model for sequence classification. In 5th IEEE International Conference on Data Mining, Houston (Texas), USA, november 2005. IEEE Computer Society.

[3] T. Minka. Discriminative models, not discriminative training. Technical report, Microsoft Research, Cambridge, UK, 2005.

[4] K.Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. International Journal of Computer Vision, 60:63–86, 2004.

[5] D. Lowe. Distinctive image features from scale invariant key points. International Journal of Computer Vision, 60(2):91–110, 2004.

[6] I. Ulusoy and C. M. Bishop. Generative versus discriminative models for object recognition. In Proceedings IEEE International Conference on Computer Vision and Pattern Recognition, CVPR, San Diego, 2005.

[7] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. Biological Cybernetics, 55:367–375, 1987.

[8] B. M. ter Haar Romay, L. M. J. Florach, A. H. Salden, and M. A. Viergever. Representation of local geometry in the visual system. Biological cybernetics, 55:367–375, 1987.

[9] G. Bouchard and B. Triggs. The trade-off between generative Computational Statistics, Pages 721– 728, Prague, Czech Republic, august 2004

[10] A.Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Advances in Neural Information Processing Systems 14, 2002

[11] C. M. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, 1995