International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:2, 2007

Density Estimation using Generalized Linear Model and a Linear Combination of Gaussians

Aly Farag, Ayman El-Baz, Refaat Mohamed

Abstract— In this paper we present a novel approach for density estimation. The proposed approach is based on using the logistic regression model to get initial density estimation for the given empirical density. The empirical data does not exactly follow the logistic regression model, so, there will be a deviation between the empirical density and the density estimated using logistic regression model. This deviation may be positive and/or negative. In this paper we use a linear combination of Gaussian (LCG) with positive and negative components as a model for this deviation. Also, we will use the expectation maximization (EM) algorithm to estimate the parameters of LCG. Experiments on real images demonstrate the accuracy of our approach.

Keywords—Logistic regression model, Expectation maximization, Segmentation.

I. INTRODUCTION

Density estimation is a core operation of virtually all probabilistic learning methods (as opposed to discriminative methods). Approaches to density estimation can be divided into two principal classes, parametric methods and nonparametric methods. While neither choice should be universally preferred for all situations, a well-known benefit of nonparametric methods is their ability to achieve good estimation for any input distribution as more data are observed, but these methods have a serious drawback, that is, there many parameters need to be tuned [1, 2].

A previous approaches for nonparametric density estimation such as the approaches proposed in [3,4] are based on the k -nearest neighbor method. It functions by saving the number of distance evaluations for finding the k nearest neighbors. As for kernel-based density estimates, such as the Parzen density estimate [2], by noting that the amount of computation is directly related to the number of training samples, Fukunaga and Hayes [5] extracted a representative subset of the training samples to achieve computational saving. The reduced subset of training samples was selected in such a way that the Parzen density estimate with the reduced set matches as closely as possible with that of the full data set in the sense of an entropy measure of similarity between the two estimates. Silverman [6] proposed an efficient algorithm based on the fast Fourier transform (FFT) for evaluating univariate Parzen density estimates on regular grids. In the case of plotting a density estimate, for example, the estimate would be evaluated over equally spaced locations. Silverman noted that the Fourier transform of the density estimate can be considered as a product of the Fourier transforms of the kernel function and the data. A modified discretization scheme was employed later by Jones and Lotwick [7] to reduce errors related to the Fourier transform of data. Notice that this fast algorithm based on FFT can not be applied to the general cases of density estimates over irregularly spaced locations. For instance, suppose Parzen density estimates are used for non-parametric classification [8], then the density values must be estimated for the samples to be classified, and in general, the samples are not regularly spaced. In [9] they proposed an approach similar to the efficient density estimate based on the k-nearest neighbors [3], they used a simple branch-and-bound procedure is applied to the Parzen density estimation to reduce the number of kernel evaluations. In [10] they proposed fast algorithm for density estimation based on using the Parzen window density estimator which employs a reduced set of the available data sample.

The basic problem in density estimation is to get accurately approximate, to within the data range, not only the peaks, or modes of the probability density function for the measurements but also its behavior between the peaks.

In this paper we introduced a new approach for density estimation that addresses the above problem. In the proposed approach we integrate the expectation maximization (EM) algorithm and logistic regression model to get accurate density estimation for the given empirical density (normalized histogram for the given data). The proposed approach consists of two main steps. First, the logistic regression model is used to get initial estimation. Then we used a linear combination of Gaussian (LCG) with positive and negative components to model this deviation. In order to estimate the parameters of this LCG (e.g weights, covariance, and means) we use EM algorithm. The final density estimation will be the summation of the density estimated using the logistic model and the error model obtained by the LCG.

Logistic regression model is a very important member of the family of generalized linear model. It became a useful tool in the 1950s in application in biostatistics.

Manuscript received November 4, 2004. This research is being supported by the United States Air Force Office of Scientific Research, AFOSR, Grant f49620-01-1-0367. The authors are with the CVIP Laboratory (www.cvip.uofl.edu), Department of Electrical and Computer Engineering, University of Louisville, Kentucky, USA. E-mail: farag@cvip.uofl.edu.

Expectation maximization algorithm is one of the most important algorithms that are used in density estimation. The algorithm appeared first in the late 1960s both in the general form [11] (see also [12]) and for a normal mixture [13]. EM algorithms became very popular a decade later, after the pivotal paper [14] extended this technique to a general problem of parameter estimation from an incomplete data set. Today a variety of EM-algorithms exist to find the maximum likelihood estimates (MLE) of parameters for mixed distributions [15-17].

II. THE PROPOSED MODEL

Given a random vector $\mathbf{Y}$, the relation:

$$P(\mathbf{Y}) = P(\mathbf{Y} < \mathbf{y})$$

defines the cumulative probability distribution function, CDF, of the random vector $\mathbf{Y}$. The probability density function, PDF, $p(\mathbf{y})$, of the random vector $\mathbf{Y}$ at a specific point $\mathbf{y}$ is a nonnegative quantity and it is related to the CDF by the relation:

$$P(\mathbf{y}) = \int_{-\infty}^{\mathbf{y}} p(\mathbf{y}')\, d\mathbf{y}' \qquad (1)$$

The density estimation problem can be stated as follows: given a random sample $D=\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n\}$ drawn from an unknown random distribution function and has empirical density function $f(\mathbf{y})$ (normalized histogram), estimate the density function defined in Eq. (1) which underlies the distribution of the sample D .

The proposed model to estimate the probability density function from any empirical density is as follows:

$$P(\mathbf{y}) = \frac{1}{1+e^{-\mathbf{y}^T\boldsymbol{\beta}}} + \zeta(\mathbf{y}) \qquad (2)$$

Where the term $\mathbf{y}^T\boldsymbol{\beta} = \beta_0 + \beta_1 y_1 + \ldots + \beta_1 y_m$ is called the **linear predictor**, and m is the number of feature in observed vector $\mathbf{y}$.

In order to make the model shown in Eq. (2) suitable also for EM algorithm, instead of estimating the cumulative probability distribution, we will estimate the probability density function. So the model of Eq. (2) can be also used for density estimation by differentiating Eq. (2) with respect to $\mathbf{y}$. So the proposed model will be as follows:

$$p(\mathbf{y}) = \frac{\boldsymbol{\beta} e^{-\mathbf{y}^T\boldsymbol{\beta}}}{1+e^{-\mathbf{y}^T\boldsymbol{\beta}}} + \varepsilon(\mathbf{y}) \qquad (3)$$

where $\varepsilon(\mathbf{y}) = \partial\xi(\mathbf{y}) / \partial\mathbf{y}$.

Now in order to find $p(\mathbf{y})$, first we need to estimate the parameters of logistic regression model $\boldsymbol{\beta}$, and to find an accurate model for the error term $\varepsilon(\mathbf{y})$. In the following sections we will explain in details how we can estimate these parameters.

A. Parameter Estimation of Logistic Regression model

In this paper, we introduced fast and accurate method to estimate the parameters of logistic regression model. This method based on finding the regression parameters that minimize the square error between the estimated density $p(\mathbf{y})$ and empirical density $f(\mathbf{y})$ as follows:

$$\boldsymbol{\beta}^* = \arg\min\left(\sum_{i=1}^{n} p(y_i) - f(y_i)\right)^2 \qquad (4)$$

To find $\boldsymbol{\beta}^*$ that minimize Eq. (4) we need nonlinear optimization approach. In this paper we used Genetic algorithm as a nonlinear global optimization approach to estimate the logistic regression parameters [18].

B. Parameter estimation for the error term $\varepsilon(\mathbf{y})$

Because the empirical density for each feature does not exactly follow the logistic model, there will be deviation between the empirical density and the estimated density. This deviation is represented in the error term $\varepsilon(\mathbf{y})$, which can be modeled by linear combination of Gaussian (LCG) with positive and negative components. The number of positive and negative components and the estimated parameters for each component is obtained using EM algorithm. The proposed model for the error term is as follows:

$$\varepsilon_{\mathbf{W},\boldsymbol{\Theta}}(\mathbf{y}) = \sum_{r=1}^{C_p} w_{p,r}\varphi(\mathbf{y}\mid\theta_{p,r}) - \sum_{l=1}^{C_n} w_{n,l}\varphi(\mathbf{y}\mid\theta_{n,l}) \qquad (5)$$

where C_p is the number of positive Gaussian components, C_n is the number of negative Gaussian components, w is the mixing proportion (weights), and $\varphi(y\mid\theta)$ is a Gaussian density having a shorthand notation $\theta = (\mu,\Sigma)$.

To estimate the parameters of the model shown in Eq. (5) we use the following algorithm

Algorithm

1. Find the deviations $\varepsilon(\mathbf{y}) = f(\mathbf{y}) - \frac{\boldsymbol{\beta} e^{-\mathbf{y}^T\boldsymbol{\beta}}}{1+e^{-\mathbf{y}^T\boldsymbol{\beta}}}$, and split them into the positive and negative parts such that $\varepsilon(\mathbf{y}) = \varepsilon_p(\mathbf{y}) - \varepsilon_n(\mathbf{y})$:
$$\varepsilon_p(\mathbf{y}) = \max\{\varepsilon(\mathbf{y}),0\}$$
$$\varepsilon_n(\mathbf{y}) = \max\{-\varepsilon(\mathbf{y}),0\}$$

.

2. Compute the scaling factor for the deviations:
$$s = \int_{-\infty}^{\infty}\varepsilon_p(\mathbf{y})d\mathbf{y} = \int_{-\infty}^{\infty}\varepsilon_p(\mathbf{y})d\mathbf{y}$$
3. If the factor s is less than an accuracy threshold, terminate and return the model $p(\mathbf{y}) = \frac{\beta e^{-\mathbf{y}^T\boldsymbol{\beta}}}{1+e^{-\mathbf{y}^T\boldsymbol{\beta}}}$

4. Otherwise consider the scaled-up absolute deviations $\frac{1}{s}\varepsilon_p(\mathbf{y})$ and $\frac{1}{s}\varepsilon_n(\mathbf{y})$ as two new "empirical densities" and use iteratively the conventional EM-algorithm [16] to find the sizes of the Gaussian mixtures and their parameters.
    a. The size of each mixture corresponds the minimum of the integral absolute error between the scaled-up absolute deviation $\varepsilon_p(\mathbf{y})$ (or $\varepsilon_n(\mathbf{y})$) and its model $p_p(\mathbf{y})$ (or $p_n(\mathbf{y})$). The number of the components is increasing sequentially by unit step while the error is decreasing.
    b. Due to multiple local maxima, such a search may be repeated several times with different initial parameter values in order to select the best approximation.
5. Scale down the subordinates models $p_p(\mathbf{y})$ and $p_n(\mathbf{y})$ (i.e. scale down the weights of their components) and add the scaled model $p_p(\mathbf{y})$ to $\frac{\boldsymbol{\beta} e^{-\mathbf{y}^T\boldsymbol{\beta}}}{1+e^{-\mathbf{y}^T\boldsymbol{\beta}}}$ and subtract the scaled model $p_n(\mathbf{y})$ form $\frac{\boldsymbol{\beta} e^{-\mathbf{y}^T\boldsymbol{\beta}}}{1+e^{-\mathbf{y}^T\boldsymbol{\beta}}}$ in order to form the desired model $p(\mathbf{y})$ as shown in Eq. (2).

We use the Levy distance [19], $\rho(P(\mathbf{y}), F(\mathbf{y}))$, between the estimated model $P(\mathbf{y})$ and the empirical distribution $F(\mathbf{y})$ to evaluate the approximation quality. The distance is defined as the minimum positive value α such that the two-sided inequalities

$$\rho(F,P) = \min_{\alpha} \{\alpha : P(\mathbf{y}-\alpha) - \alpha \le F(\mathbf{y}) \le P(\mathbf{y}+\alpha) + \alpha \ \forall \mathbf{y} \in \mathbf{Y}\} \tag{6}$$

III. EXPERIMENTAL RESULT AND CONCLUSIONS

The application domain in this case is medical screening of the bi-modal LDCT images having two dominant regions: the darker lungs and the brighter background consisting of the surrounding anatomical structures (e.g., chest, ribs, and liver). The lung tissues have to be accurately separated from the background in such a way that the lung borders closely approach the borders outlined by a radiologist. Figure 1 shows a typical CT slice (a) and its empirical density for both the lung and chest tissues (b). Note that these two empirical densities are calculated from 75% of the given data and the remaining 25% will be used for testing. Figure 2(a) shows the initial estimated logistic model for the empirical density of the lung tissues shown in Fig. 1(b). In this example we use only one feature (gray level intensity) so the logistic model has only two parameters $(\boldsymbol{\beta} = [\beta_0 \quad \beta_1])$. These two parameters are estimated using genetic algorithm and their values are shown in Table 1. The Levy distance of 0.16 between these two distributions (see Eq. (6)) indicates a large mismatch between the empirical distribution and the initial density estimation obtained by logistic model. Figure 2(b) shows the scaled absolute deviation between the estimated density using the logistic regression model and the empirical density of lung tissues and its estimated model using LCG. Figure 2(c) shows the components used for density estimation (logistic model, and LCG components). The final estimated density for the lung empirical density is shown in Fig. 2(d). The resulting Levy distance of 0.0094 is notably smaller than before (0.16) indicating the close approximation. In the same way we will follow the same procedure to estimate the density for the chest tissues. The results for density estimation for the chest tissues is shown in Fig.3 and Table 1. Figure 4(a) shows the result of lung segmentation obtained by using Bayes classifier with equal a priori probability. The proposed approach achieves segmentation with accuracy 3.1% with respect to radiologist segmentation shown in Fig. 4(b). Figure 5 shows more results obtained by the proposed approach.

These results and other experiments with multimodal medical images and remote sensing data show that the proposed approach produces accurate density estimation if it is compared with other existing approaches and this can help in many applications in image processing field such as image segmentation. The computations of the proposed approach is so simple and it takes few seconds for estimating any density (e.g. the algorithm takes 21 second to estimate the density for the lung tissues)

REFERENCES

[1] V. N. Vapnik, Statistical Learning Theory. New York: John Wiley and Sons, 1998.
[2] B.W. Silverman, Density Estimation for Statistics and Data Analysis. Chapman and Hall, 1986.
[3] P. E. Hart, "The Condensed Nearest Neighbor Rule*," IEEE Transactions on. Information Theory*, IT-14, pp.515-516, 1968.
[4] K. Fukunaga and P. M. Narendra, "A Branch And Bound Algorithm For Computing K-Nearest Neighbors," *IEEE Transactions on Computers*, C-24, pp. 750-753, 1975.
[5] K. Fukunaga and R. R. Hayes, "The Reduced Parzen Classifier," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11, pp.423-425, 1989.
[6] B. W. Silverman, "Kernel Density Estimation Using The Fast Fourier Transform," *Statistical Algorithm, AS176, Applied Statistic*, 31, pp.93-97, 1982.
[7] M. C. Jones and H. W. Lotwick, "A Remark on Algorithm AS 176. Kernel Density Estimation Using the Fast Fourier Transform," *Remark, AS R50, Applied Statistic*, 33, pp.120-122, 1984.
[8] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press, New York, 1990.
[9] B. Jeon; Landgrebe, D.A., "Fast Parzen density estimation using clustering-based branch and bound," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16 , Issue: 9 September, 1994 pp. 950 – 954.
[10] M. Girolami and C. He, "Probability Density Estimation from Optimally Condensed Data Samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, No. 10, October 2003, pp. 1253 – 1264.
[11] M. I. Schlesinger and V. Hlavac, "Ten Lectures on Statistical and Structural Pattern Recognition" Kluwer Academic: Dordrecht, 2002.

[12] M. I. Schlesinger, "A connection between supervised and unsupervised learning in pattern recognition", Kibernetika, no.2, pp. 81–88, 1968 [In Russian].

[13] N. E. Day, "Estimating the components of mixture of normal distributions", Biometrika, Vol. 56, pp. 463–474, 1969.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistics Society*, vol. 39B, no. 1, pp. 1–38, 1977.

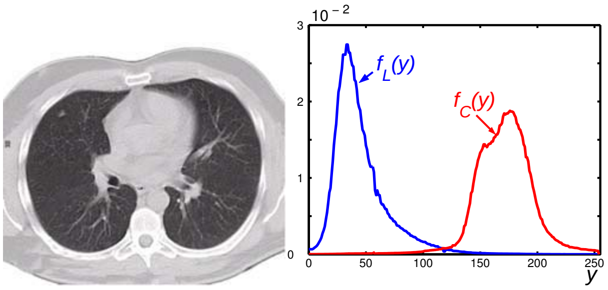
[15] C. J. McLachlan, The EM algorithm and extensions, Wiley, New York, 1997.

[16] T. Moon, "The Expectation - Maximization algorithm", *IEEE Signal Processing Magazine*, no. 11, pp. 47–60, 1996.

[17] R. Redner and H. Walker, "Mixture densities, maximum likelihood and the EM algorithm (review)", *SIAM Review*, Vol. 26, no. 2, pp. 195–237, 1984.

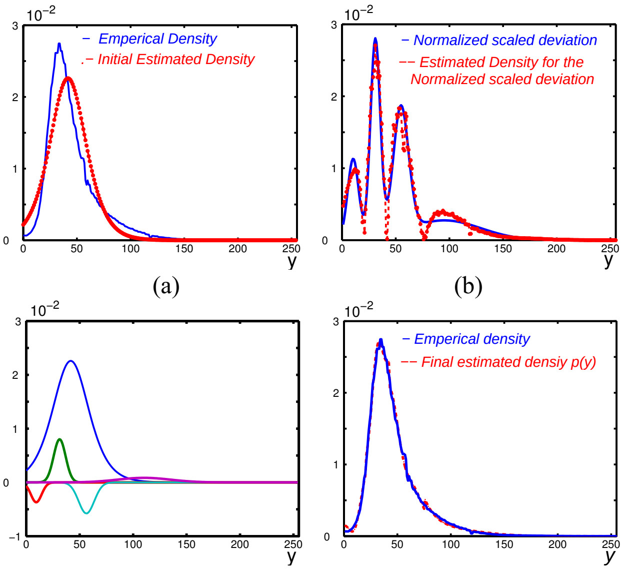
[18] D.E.Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Boston, 1989.

[19] J.W.Lamperti, Probability, J. Wiley and Sons, New York, 1996.



(a) (b)

Fig. 1 (a) Typical CT slices, (b) Empirical density for both the lung tissues $f_L(\mathbf{y})$, and the chest tissues $f_C(\mathbf{y})$.



(a) (b)

(c) (d)

Fig.2 (a) Initial estimated density for the lung tissues, (b) Estimated density for the absolute scaled deviation, (c) All components used for estimating the intensity of lung tissues, (d) Final density estimation for the lung tissues.

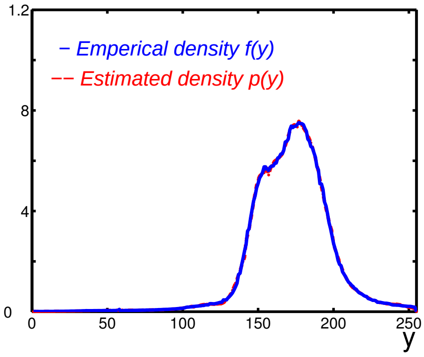


Fig. 3 Final density estimation for the lung tissues using the proposed approach.

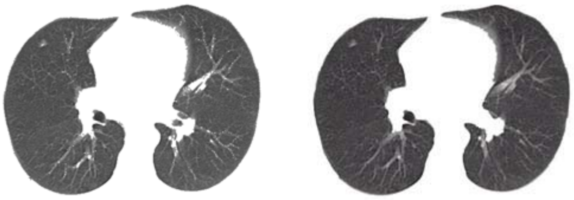
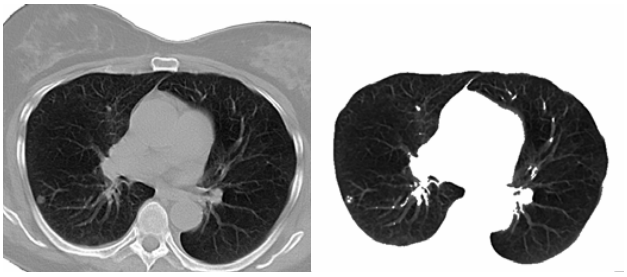


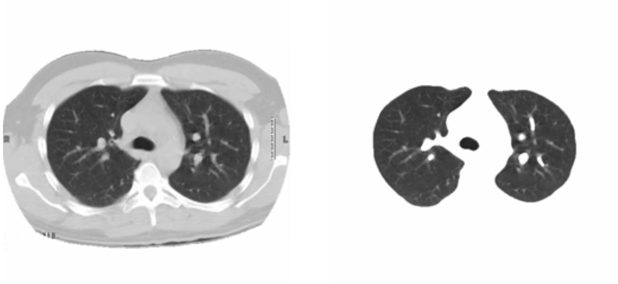
Fig.4 (a) Segmentation by the proposed approach (Error 3.1%), (b) Ground truth generated by radiologist.

Table 1. Estimated parameters for logistic regression model for both lung and chest empirical densities.

Parameters	Lung tissues	Chest tissues
β_0	-4.1	-12.1
β_1	1.05	0.21



Error = 1.9%



Error = 1.7%

Fig. 5 (a) The original images, (b) The segmentation obtained by the proposed approach