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

# Probability Density Estimation Using Advanced Support Vector Machines and the Expectation Maximization Algorithm

Refaat M Mohamed, Ayman El-Baz and Aly A. Farag, Senior Member IEEE *

***Abstract*** This paper presents a new approach for the probability density function estimation using the Support Vector Machines (SVM) and the Expectation Maximization (EM) algorithms. In the proposed approach, an advanced algorithm for the SVM density estimation which incorporates the Mean Field theory in the learning process is used. Instead of using ad-hoc values for the parameters of the kernel function which is used by the SVM algorithm, the proposed approach uses the EM algorithm for an automatic optimization of the kernel. Experimental evaluation using simulated data set shows encouraging results.

***Keywords*** Density Estimation, SVM, Learning Algorithms, Parameters Estimation.

## I. INTRODUCTION

DENSITY estimation is a major ingredient in Bayesian pattern recognition and machine learning. Many algorithms have been introduced for solving the density estimation problem [1]. Support Vector Machines (SVM) have been developed by Vapnik [2] to solve the classification problem, but recently, SVM have been successfully extended to regression and density estimation problems [3]. SVM are gaining popularity due to many attractive features and promising empirical performance. For instance, the formulation of SVM density estimation embodies the Structural Risk Minimization (SRM) principle which has been shown to be superior to traditional Empirical Risk Minimization (ERM) principle employed in conventional learning algorithms (e.g. neural networks) [4]. It is this difference which makes SVM more attractive in statistical learning applications.

The traditional formulation of the SVM density estimation problem raises a Quadratic optimization Problem (QP) of the same size as the training data set. The (QP) is computationally expensive and solving such a problem is not trivial [5].

In this paper, a new formulation of the SVM density estimation problem is proposed. This formulation enables the use of the Mean Field theory (MF) in the learning of the SVM algorithm. The MF methods provide efficient approximations which are able to cope with the complexity of probabilistic data models, see [6]. MF methods replace the intractable task of computing high dimensional sums and integrals by the much easier problem of solving a system of linear equations.

The Expectation Maximization (EM) algorithm is a general method of finding the maximum-likelihood estimate (MLE) of the parameters of an underlying distribution from a given data set, when the data set is incomplete or has missing values. In this paper, the EM algorithm is used for automatic selection of the kernel function parameters instead of using ad-hoc values for these parameters. As a common choice, a Gaussian Radial Basis(GRB) kernel is used in the proposed approach. The EM algorithm is used to automatically select the covariance matrices of the kernels centered at the training instants.

Experimental evaluation of the proposed algorithm is carried out using synthetic data. The performance of the proposed algorithm is compared with previous algorithms. The evaluation is based on the Levy distance and the Kullback-Leibler distance. The results show encouraging performance of the proposed SVM density estimation algorithm with automatic selection of the kernel parameters.

*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.

## II. DENSITY ESTIMATION USING THE SVM

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), $p(\mathbf{y})$, is a nonnegative quantity and it is related to the CDF by the relation: $P(\mathbf{y}) = \int_{-\infty}^{\mathbf{y}} p(\mathbf{y}')\, d\mathbf{y}'$. The density estimation problem can be stated as follows: given a random sample $\mathbf{D} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ drawn from an unknown random distribution, estimate the density function $p(\mathbf{y})$ which underlies the distribution of the sample $\mathbf{D}$.

From the above discussion, the probability density function $p(\mathbf{y})$ can be easily obtained from:

$$p(\mathbf{y}) = \frac{d}{d\mathbf{y}} P(\mathbf{y}) \tag{1}$$

The empirical cumulative distribution function $P_n(\mathbf{y})$ is defined as:

$$P_n(\mathbf{y}) = \frac{1}{n} \sum_{k=1}^{n} I_{(-\infty,\, \mathbf{y}]}(\mathbf{y}_k) \tag{2}$$

where, $I_{(-\infty,\, \mathbf{y}]}(u)$ is the indicator function. $P_n(\mathbf{y})$ converges in probability to the cumulative distribution function $P(\mathbf{y})$ (e.g., [7] ).

Thus, one of the suggested solutions of the density estimation problem is to use $P_n(\mathbf{y})$ to get an approximation for $P(\mathbf{y})$. Then, the approximate $P(\mathbf{y})$ is differentiated to get $p(\mathbf{y})$ (see Eq.(1)). So, a training data set $\mathcal{D} = \{(\mathbf{y}_i, t_i) :\ t_i = P_n(\mathbf{y}_i); i = 1, 2, \ldots, n\}$ is generated from $\mathbf{D}$ and it is used by a regression algorithm to solve the density estimation problem in Eq. (1), in the data (image) space, in order to obtain a continuous approximation for the distribution function. This approximation can be used to express the solution in the pre-image space to get an estimate for the density function.

The Support Vector Machines (SVM) can be employed to obtain a continuous approximation for the distribution function. The motivation behind using the SVM as a regression tool is that a dense continuously differentiable approximation for the distribution function can be obtained which can be safely differentiated to obtain the density function.

### A. SVM Regression

In this section the Support Vector Machines (SVM) as a regression tool is presented. The detail foundation of the SVM regression can be found in our previous work [3], but here only the main outlines of the algorithm are presented. In the following, the SVM as a regression tool is considered as the maximum a posteriori (MAP) prediction with a Gaussian prior under the Bayesian framework. Thus, the output from the SVM regression for the sample $\mathcal{D}$ is represented as a Gaussian process with a zero mean in the following form:

$$\begin{aligned} p(\mathbf{g}(\mathcal{D})) &\sim \mathcal{N}\left(\mathbf{g}\left(\mathcal{D}\right)|\mathbf{0},\mathcal{K}_n\right) \\ &= \frac{1}{\sqrt{(2\pi)^n\, det\left(\mathcal{K}_n\right)}} \exp\left\{-\frac{1}{2}\mathbf{g}\left(\mathcal{D}\right)\mathcal{K}_n^{-1}\mathbf{g}\left(\mathcal{D}\right)^T\right\} \end{aligned} \tag{3}$$

where $\mathcal{K}_n = [\mathcal{K}(\mathbf{y}_i,\mathbf{y}_j)]$ is the covariance matrix at the points of $\mathcal{D}$ and $\mathbf{g}\left(\mathcal{D}\right) = [g\left(\mathbf{y}_i\right)]$ is the SVM output vector.

The performance of the SVM regression algorithm is characterized by the Vapnik's $\varepsilon$-loss function which has the form:

$$\mathcal{L}\left(t, g(\mathbf{y})\right) = \begin{cases} 0 & \text{if } |t-g(\mathbf{y})| \le \epsilon \\ |t-g(\mathbf{y})| - \epsilon & \text{otherwise} \end{cases} \tag{4}$$

Depending on this loss function, the likelihood of the target output vector $\mathcal{T}$ given the actual SVM output will be in the form:

$$p(\mathcal{T}|\mathbf{g}(\mathcal{D}))= \left(\frac{C}{2\left(\epsilon C+1\right)}\right)^n \exp\left\{-C\sum_{i=1}^{n}\mathcal{L}(t_i, g(\mathbf{y}_i))\right\} \tag{5}$$

where: $\mathcal{T}= [t_1, t_2, \ldots, t_n]$.

Using Equations (3) and (5) and from Bayes' theorem:

$$\begin{aligned} p\left(\mathbf{g}\left(\mathcal{D}\right)|\mathcal{D}\right) &= \frac{p\left(\mathcal{D}|\mathbf{g}\left(\mathcal{D}\right)\right)p\left(\mathbf{g}\left(\mathcal{D}\right)\right)}{p\left(\mathcal{D}\right)} \\ &= \frac{\mathcal{M}\exp\left\{-C\sum_{i=1}^{n}\mathcal{L}(t_i, g(\mathbf{y}_i)) - \frac{1}{2}\mathbf{g}\left(\mathcal{D}\right)\mathcal{K}_n^{-1}\mathbf{g}\left(\mathcal{D}\right)^T\right\}}{\sqrt{2\pi^n\, det\left(\mathcal{K}_n\right)}\; p\left(\mathcal{D}\right)} \end{aligned} \tag{6}$$

where $\mathcal{M} = \left(\frac{C}{2(\epsilon C+1)}\right)^n$.

Using the posterior prediction distribution $p\left(\mathbf{g}\left(\mathcal{D}\right)|\mathcal{D}\right)$ which is defined in Eq.(6), the predicted (expected) SVM output on a new test point $\mathbf{y}$ is given by:

$$\begin{aligned} \langle g(\mathbf{y})\rangle &= \int g\left(\mathbf{y}\right) p\left(g\left(\mathbf{y}\right)|\mathcal{D}\right)\, dg(\mathbf{y}) \\ &= \int g\left(\mathbf{y}\right) p\left(g\left(\mathbf{y}\right), \mathbf{g}\left(\mathcal{D}\right)|\mathcal{D}\right)\, dg(\mathbf{y})\; d\,\mathbf{g}(\mathcal{D}) \end{aligned} \tag{7}$$

Substituting in Eq.( 7) from the previous equations and after some mathematical reductions, the output $g(\mathbf{y})$ of the SVM regression algorithm will have the form:

$$\langle g(\mathbf{y})\rangle = \sum_{i=1}^{n} w_i\, \mathcal{K}\left(\mathbf{y}, \mathbf{y}_i\right) \tag{8}$$

where $\mathcal{K}(\mathbf{y}, \mathbf{y}_i)$ is the kernel function used by the SVM regression algorithm and $w_i$'s are the weight coefficients where $w_i$ is given by:

$$\begin{aligned} w_i &= \frac{\mathcal{M}}{p\left(\mathcal{D}\right)}\int \mathcal{N}\left(\mathbf{g}\left(\mathcal{D}\right)|\mathbf{0},\mathcal{K}_n\right)\left(\mathbf{g}\left(\mathcal{D}\right)|\mathbf{0},\mathcal{K}_n\right) g\left(\mathbf{y}\right). \\ &\quad \frac{\partial}{\partial g\left(\mathbf{y}_i\right)}\exp\left\{-C\sum_{j=1}^{n}\mathcal{L}(t_j, g(\mathbf{y}_j))\right\} d\mathbf{g}(\mathcal{D}) \end{aligned} \tag{9}$$

The learning process suggests that the weights $w_i$'s would be estimated using the training data set $\mathcal{D}$. One way to enable this estimation is to consider the $i$'th sample from $\mathcal{D}$ as a test sample and to estimate the corresponding $i$'th weight $w_i$ using the rest of the training data set $\overline{\mathcal{D}} = \{\mathcal{D}\} - \{(\mathbf{y}_i, t_i)\}$. This estimation requires a definition for the predictive distribution $p(g(\mathbf{y}_i)|\overline{\mathcal{D}})$, which is called the *cavity distribution*. Suppose that the cavity distribution is expressed in the form:

$$p\left(g\left(\mathbf{y}_i\right)|\overline{\mathcal{D}}\right) = \frac{\mathcal{N}(\mathbf{g}(\mathcal{D})|\mathbf{0},\mathcal{K}_n)\exp\left\{-C\sum_{i\neq j}\mathcal{L}(t_j - g(\mathbf{y}_j))\right\} d\mathbf{g}\left(\overline{\mathcal{D}}\right)}{\mathcal{N}(\mathbf{g}(\mathcal{D})|\mathbf{0},\mathcal{K}_n)\exp\left\{-C\sum_{i\neq j}\mathcal{L}(t_j - g(\mathbf{y}_j))\right\} d\mathbf{g}(\mathcal{D})} \tag{10}$$

Then the weight $w_i$ in Eq.(9) can be expressed as:

$$w_i = \frac{\langle \mathcal{M}\frac{\partial}{\partial g(\mathbf{y}_i)}\exp\left\{-C\mathcal{L}(t_j, g(\mathbf{y}_j))\right\}\rangle_i}{\langle \mathcal{M}\exp\left\{-C\mathcal{L}(t_j, g(\mathbf{y}_j))\right\}\rangle_i} \tag{11}$$

where $\langle\nu\rangle_i$ denotes the expected value of $\nu$ with respect to the cavity distribution $p(g(\mathbf{y}_i)|\overline{\mathcal{D}})$.

### B. Mean Field Theory for SVM Learning

Obtaining the weights using the formula in Eq.(11) is intractable due to the complicated vector integrations which in turn need highly computationally expensive numerical integration methods. One of the new algorithms to approximate the estimation of these weights is to use the Mean Field theory. The basic idea of the Mean Field theory is to approximate the statistics of a random variable which is correlated to other random variables by assuming that the influence of the other variables can be compressed into a single effective mean "field" with a rather simple distribution. In this paper, the principle of the Mean Field theory is used to approximate the cavity distribution $p\left(g\left(\mathbf{y}_i\right)|\overline{\mathcal{D}}\right)$. The approximation here is carried out by assuming a simple form (a Gaussian distribution is used in this paper) for the cavity distribution which enables the calculation of the weights. Depending on the assumed form for $p\left(g\left(\mathbf{y}_i\right)|\overline{\mathcal{D}}\right)$, the weight $w_i$ can be obtained from Eq.( 11). The details for the learning procedure can be found in [3].

### C. Obtaining the Density Function Estimate

The above discussion shows how the SVM can be used as a regression tool. In this paper, the SVM regression algorithm is used to approximate the distribution function $P(\mathbf{y})$ from the training sample $\mathcal{D}$. The approximation will be in the form of a weighted sum of the kernel function working on the instants of the training sample as:$P\left(\mathbf{y}\right) = \sum_{i=1}^{n} w_i\, \mathcal{K}\left(\mathbf{y}, \mathbf{y}_i\right)$. Consequently, the estimate of the density function will be simply in the form:

$$p\left(\mathbf{y}\right) = \sum_{i=1}^{n} w_i\, \mathcal{K}'\left(\mathbf{y}, \mathbf{y}_i\right) = \sum_{i=1}^{n} w_i\, \mathfrak{K}\left(\mathbf{y}, \mathbf{y}_i\right) \tag{12}$$

where $\mathfrak{K}\left(\mathbf{y}, \mathbf{y}_i\right)$ is the derivative of $\mathcal{K}\left(\mathbf{y}, \mathbf{y}_i\right)$.

There are some conditions on the kernel function $\mathfrak{K}\left(\mathbf{y}, \mathbf{y}_i\right)$ so that a valid density function estimate can be obtained from Eq.(12), see [2]. These conditions are:
$\mathfrak{K}_\gamma = a\left(\gamma\right)\mathfrak{K}\left(\frac{\mathbf{y}-\mathbf{y}_i}{\gamma}\right)$, $a\left(\gamma\right)\int \mathfrak{K}\left(\frac{\mathbf{y}-\mathbf{y}_i}{\gamma}\right) d\mathbf{y} = 1$ and $\mathfrak{K}(0) = 1$.

## III. Optimization of the Kernel Function

One of the commonly used kernels is the Gaussian Radial Basis Function (GRBF) which satisfies the above conditions and it has

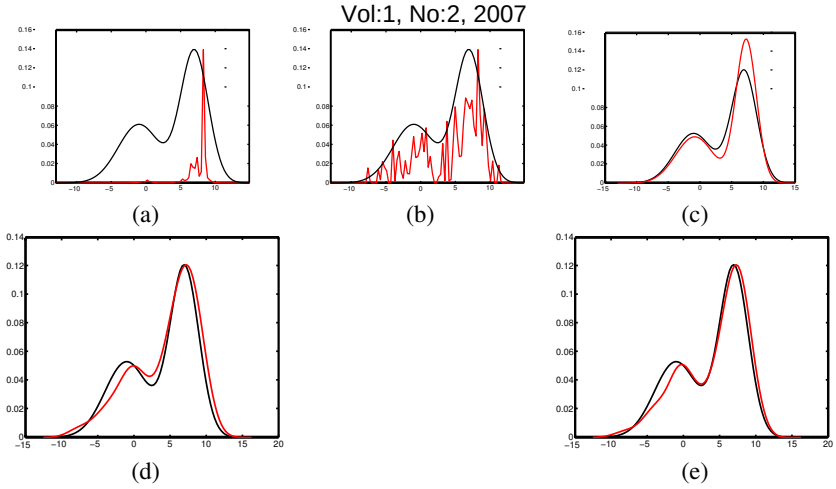


Figure 1: Estimation of a mixture of Gaussians density function with a)KNN, (b) Parzen-window, (c) EM, (d) MF-SVM without automatic kernel optimization, and (e) proposed algorithm. (Black:reference, Red:estimated)

the form:

$$\Re(\mathbf{y}, \mathbf{y}_i) = \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{y}_i)\Lambda^{-1}(\mathbf{y}-\mathbf{y}_i)^T\right) \quad (13)$$

where $\Lambda$ is the covariance which is selected empirically in the traditional methods (see [4]).

In this paper, an approach for automatic selection of the covariance for the RBF kernel is suggested. This approach incorporates the EM algorithm into the learning procedure so that the covariance of the kernel is optimized while the SVM weight coefficients are estimated.

The EM algorithm can be used to estimate the parameters of a mixture of a Gaussian distribution based on the maximization of the following likelihood function:

$$L(\mathbf{w}, \mathbf{\Theta}) = \sum_{\mathbf{y}\in\mathbf{Y}} f(\mathbf{y}) \log p((\mathbf{y})) \quad (14)$$

where $f(\mathbf{y})$ is the empirical density function.

The maximization of Eq.(14) can be found using the iterative block relaxation algorithm. The relative contributions of each data item $\mathbf{y}$ into each Gaussian component at the step $m$ are specified by the following respective conditional weights

$$\pi^{[m]}(r|\mathbf{y}) = \frac{w_r^{[m]}\varphi(\mathbf{y}|\theta_r^{[m]})}{p^{[m]}_{\mathbf{w},\mathbf{\Theta}}(\mathbf{y})} \quad (15)$$

where $r = 1, 2, \ldots, n$. The block relaxation converging to a local maximum of the likelihood function in Eq. (14). The following two steps are repeated iteratively to get the parameters of the mixture:

1. E-step $[m+1]$: to find the covariance of a Gaussian component by maximizing $L(\mathbf{w}, \mathbf{\Theta})$ under the fixed conditional weights of Eq. (15) for the step $m$, and

2. M-step $[m+1]$: to find these latter weights by maximizing $L(\mathbf{w}, \mathbf{\Theta})$ under the fixed parameters (in our case is the covariance)

until the changes of the log-likelihood and all the model parameters become small. The covariance of each Gaussian are obtained by the unconditional maximization:

$$(\sigma_r^{[m+1]})^2 = \frac{1}{w_r^{[m+1]}} \sum_{\mathbf{y}\in\mathbf{Y}} \left(\mathbf{y}-\mu_i^{[m+1]}\right)\cdot\left(\mathbf{y}-\mu_i^{[m+1]}\right)' \cdot f(\mathbf{y})\pi^{[m]}(r|\mathbf{y}) \quad (16)$$

## IV. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm, a data set of 100 instants is generated from a 1-D mixture of Gaussians. The mixture consists of two components and has the form:

$$p(\mathbf{x}) = \alpha_1\,\varphi(\mu_1, \sigma_1^2) + \alpha_2\,\varphi(\mu_2, \sigma_2^2) \quad (17)$$

with the parameters shown in Table 1.

Table 1: Parameters of the 1-D mixture of Gaussians

| **Parameter** | $\mu_1$ | $\mu_2$ | $\sigma_1^2$ | $\sigma_2^2$ | $\alpha_1$ | $\alpha_2$ |
|---|---|---|---|---|---|---|
| **Value** | -1 | 7 | 9 | 4 | 0.6 | 0.4 |

The results in Fig. (1) show that the proposed Mean Field-based SVM density estimation with automatic kernel optimization approach approximates well the density function in Eq. (17). It outperforms all other algorithms either classical algorithms (e.g. KNN, Parzen-window and EM) or new algorithms (e.g. MF-SVM without automatic kernel optimization). For a quantitative evaluation, the Kullback-Leibler distance (KLD) and the Levy distance measures are used.

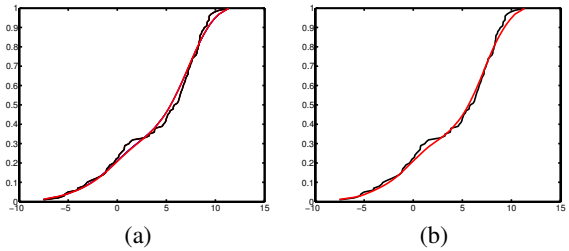
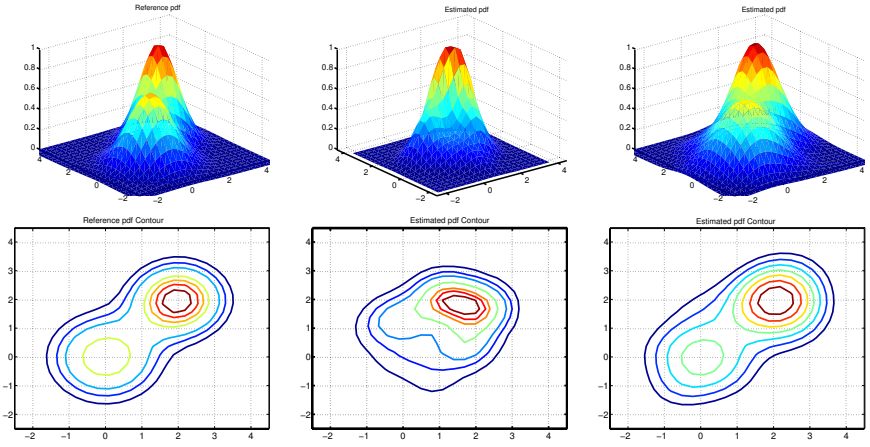


Figure 2: CDF of the densities estimated by (a) MF-SVM without automatic kernel optimization, and (b) proposed algorithm.(Black:empirical, Red:estimated)




(a) (b) (c)

Figure 3: Estimation of a 2-D Gaussian density function, (a) reference and its contour, (b) estimated using traditional formulation and its contour, (c)estimated using MF-based SVM and its contour.

The KLD is used to quantitatively compare two density functions and it is known that the KLD tends to zero for a good density estimator. For the proposed MF-based SVM approach with kernel optimization, the KLD is 0.02 which is small enough to show that the proposed approach is a good density estimator. For comparison purposes, the KLD for the nearest performing approach (see Fig.(1)) which is the MF-Based SVM approach without kernel optimization is 0.09, which is another figure for the outperforming of the proposed approach over other algorithms.

The Levy distance is used to compare two distribution functions in order to reflect the similarity of their density functions. In this experiment, the Levy distance is used to compare the empirical distribution function of the input random sample and the estimated distribution function by the density estimator. The CDF of the MF-Based SVM without kernel optimization and that of the proposed MF-based SVM with kernel optimization are shown in Fig. (2). The Levy distance is 0.049 for the proposed approach while it is 0.079 for the MF-Based SVM without kernel optimization which again illustrates the outstanding performance of the proposed approach.

Another experiment is carried out to illustrate the performance of the proposed algorithm in higher dimensional spaces. A data set is generated with 100 instants from a 2-D mixture of Gaussians distribution with the parameters in Table 2.

Table 2: Parameters of the 2-D mixture of Gaussians

| **Parameter** | $\mu_1$ | $\mu_2$ | $\Sigma_1^2$ | $\Sigma_2^2$ | $\alpha_1$ | $\alpha_2$ |
|---|---|---|---|---|---|---|
| **Value** | $\begin{bmatrix}0\\0\end{bmatrix}$ | $\begin{bmatrix}2\\2\end{bmatrix}$ | $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ | $\begin{bmatrix}0.6 & 0\\0 & 0.6\end{bmatrix}$ | 0.6 | 0.4 |

This experiment is used to compare the proposed algorithm with the traditionally formulated SVM algorithm. Fig. (3) shows both the density function and its contour for the reference density function, the estimated density function using the traditionally-formulated SVM estimator, and the estimated density function using the proposed MF-Based SVM estimator. As illustrated visually by the figure, there is a significant improvement of the performance using the MF-based SVM over the traditionally formulated SVM. In the contour plot for the estimated density there is an apparent deformation in the contour of the estimated density function using the traditional SVM where the KLD in this case is 8.4 which is quite high indicating a poor estimation. In the MF-Based SVM, the contour plot of the stimated ensity is close to the contour of the referencee one where the KLD is 0.1 which emphasizes the good performance of the proposed algorithm.

## V. CONCLUSION

An approach for solving the density function estimation problem is presented. This approach incorporates the Mean Field theory in the learning of the SVM to overcome the problems of the traditional formulation of the SVM learning. An automatic method for selecting the parameters of the SVM kernel is also proposed. This automatic method incorporates the EM algorithm in the learning procedure of the MF-based SVM. The experimental results show that the performance of the proposed approach outperforms other algorithms either classical or new algorithms.

## References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification. 2nd ed., Wiley: New York, 2000.

[2] V. Vapnik, The Nature of Statistical Learning Theory. 2nd ed., Springer: New York, 2001.

[3] Refaat M. Mohamed and Aly A. Farag, "Mean Field Theory for Density Estimation Using Support Vector Machines," Seventh International Conference on Information Fusion, Stockholm, July, 2004, pp. 495-501.

[4] V. Vapnik, S. Golowich and A. Smola, "Support Vector Method for Multivariate Density Estimation," Proc. Neural Information Processing Systems 1999 (NIPS 99), Vol. 9, MIT Press:Cambridge, MA, 2000.

[5] B. Scholkopf , C. Burges, and A. Smola, Advances in Kernel Methods:Support Vector Learning. MIT Press:Cambridge, MA, 1999.

[6] Manfred Opper and Ole Winther, "Gaussian Processes for Classification: Mean Field Algorithms," Neural Computation., Vol. 12, pp. 2655–2684, 2000.

[7] John W. Lamperti, Probability-A survey of the Mathematical Theory. Wiley Series in probability and Statistics, New York, 1996.