# On Bayesian Inference, Maximum Entropy and Support Vector Machines Methods

Mihai Costache*, Marie Liénou* and Mihai Datcu†,*

*GET-Télécom Paris - 46 rue Barrault, 75013 Paris, France
†German Aerospace Center DLR - Oberpfaffenhofen, D-82234 Wessling, Germany

**Abstract.** The analysis of discrimination, feature and model selection conduct to the discussion of the relationships between Support Vector Machine (SVM), Bayesian and Maximum Entropy (MaxEnt) formalisms. MaxEnt discrimination can be seen as a particular case of Bayesian inference, which at its turn can be seen as a regularization approach applicable to SVM. Probability measures can be attached to each feature vector, thus feature selection can be described by a discriminative model over the feature space. Further the probabilistic SVM allows to define a posterior probability model for a classifier. In addition, the similarities with the kernels based on Kullback-Leibler divergence can be deduced, thus returning with MaxEnt similarity.

**Keywords:** Perceptron, Support Vector Machine, Bayesian Inference, Maximum Entropy.
**PACS:** 89.20.Ff

## INTRODUCTION

The analysis of discrimination, feature and model selection -conduct to the discussion of the relationships between different classification formalisms based on machine learning methods, such as Support Vector Machine (SVM), Bayesian and Maximum Entropy (MaxEnt) inference. Therefore, each of the methods can be linked with the others in a particular manner and is characterised by similarities and particularities.

The present article illustrates these connections with an incursion in the history of the classification formalisms, taking into consideration the evolution from the original simple linear classifier, the Perceptron, up to the Maximum Entropy formalism.

Each classification method is using a decision function $f$ whose parameters are determined in the training stage and then used for classification purposes. Comparison between different formalisms will take into consideration the similarities and particularities concerning the two steps. In order to keep it as clear and simple as possible, one will define the learning problem in the case of the binary classification. Thus the task is to find the decision function $f$ which, based on independent observations $D$, assigns an instance $x$ of the data $D$ to one of the two classes denoted by $y \in \{+1,-1\}$.

The general form of the decision function $f$ is given by:

$$f(x) = sgn(g(x)) \quad and \quad g(x) = (w \cdot x + b) \tag{1}$$

where '$\cdot$' represents the dot product and $w$ and $b$ are the parameters to be determined. The sign of $f(x)$ is used to classify the input data $x$ into two classes. The considered formalisms have different approaches. We will be interested in the relations between each of them and under what conditions a formalism can be seen as a particular case of another.

This paper is organised as follows: the next section describes the Perceptron, SVM and Radial Basis Function (RBF) formalisms, then the third section is dedicated to the Bayesian approach and its connections with the already introduced SVM and RBF methods. Section four is introducing the MaxEnt formalism which can be seen as a particular case of Bayesian approach, and its relations with SVM by means of kernel function. In the last section, are discussed practical aspects of SVM and Bayesian methods in Information Retrieval (IR).

## SUPPORT VECTOR MACHINES

The SVM principle has his roots in the well known Perceptron formalism. The Perceptron is one of the first binary linear classifiers and represents the simplest kind of feedfoward Neural Networks. The principle is simple: an input vector *x* is transposed into an output value of the decision function *f(x)* given by Eq. 1. The problem to solve in order to perform classification is to determine the weight vector *w* and the scalar *b* starting from a training sequence. Different algorithms can be employed for this purpose: Stochastic Gradient Descent, Mean Square Error and Cross-Entropy [3].

### The SVM formalism

In the last years, much attention was devoted to the powerful kernel-based learning SVM formalism. The kernel-based machine learning algorithms are used for data which are not linearly separable. For this reason a function $\Phi(x)$ maps the data into a new highly dimensional space where the classification task is linear. The main idea in the SVM formalism is, based on the training set, to trace two surfaces that best delimitate the examples in two classes so that the area between them, called margin area, be maximised with minimum of training error. The instances which are on the two delimitation surfaces are called Support Vectors (SV) and they are used in the classification step. Having the decision function $f$ as $f(x) = sgn(w \cdot \Phi(x) + b)$, one can express the condition of perfect classification, taking into consideration the observed data used for training step, as

$$y_i((w \cdot \Phi(x_i) + b) \geq 1 \quad i = 1, \cdots, n \tag{2}$$

with $y_i$ representing the labels of the instances. In order to meet such conditions, one has to minimise the expected risk, thus $||w||^2$, as stated in [4]. As the margin area is given by $\frac{2}{||w||}$, the problem is translated into a margin area maximisation. Introducing the Lagrange multipliers $\alpha_i$, $i = 1, \cdots, n$ for each of the conditions in Eq. 2, one obtains:

$$W(\alpha) = \frac{1}{2}||w||^2 - \sum_{i}^{n} \alpha_i (y_i(w \cdot \Phi(x_i) + b) - 1). \tag{3}$$

Minimising Eq. 3 with respect to $w$ and $b$ and maximising it with respect to the Lagrangian multipliers gives $\sum_{i=1}^{i=n} \alpha_i y_i = 0$ and $w = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i)$. In this way, the problem changes to a quadratic optimisation one:

$$\max_{\alpha} \Big( \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\Phi(x_i) \cdot \Phi(x_j)) \Big) \tag{4}$$

subject to $\alpha_i \geq 0, i = 1, \cdots, n$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. The dot product from Eq. 4 is replaced with a kernel function $K$:

$$K(x_i, x_j) = < \Phi(x_i), \Phi(x_j) > \tag{5}$$

In this way, the decision task in the new vectorial space can be solved with no need of knowledge over the mapping function $\Phi(x)$. Having solved the dual optimisation problem given by Eq. 4, the Lagrange multipliers $\alpha_i$, $i = 1, \cdots, n$ are obtained and used to compute the decision function as follows:

$$f(x) = sgn(g(x)) \quad and \quad g(x) = (\sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b). \tag{6}$$

Taking into consideration that the real case data are noisy, and in order to prevent overfitting, slack variables are introduced to relax the hard margins constraints in Eq. 2, which becomes $y_i((w \cdot \Phi(x_i) + b) > 1 - \varepsilon_i, \quad \varepsilon_i \leq 0, i = 1, \cdots, n$.

Based on the above observations, one can state that the Perceptron is equivalent to the linear SVM, with the only difference appearing for the training procedure. Indeed, in the case of Perceptron, the instances are linearly separable and only one separation surface is determined while in the SVM approach, two separation surfaces are needed.

Based on the Perceptron, more complex methods have been derived, such as Neural Networks. There are many types of Neural Networks, each of them with its own particularities. Among them, the particular case of RBF is considered in this paper.

## Radial Basis Function

A special case of the Neural Networks is the Radial Basis Function (RBF). It consists of a classifier for which the decision function $f$ can be written as follows:

$$f(x) = sgn(g(x)) \quad and \quad g(x) = (\sum_{i=1}^{n} w_i \cdot exp(-\frac{||x - x_i||^2}{c_i}) + b) \tag{7}$$

with $x_i$ representing the centre and $c_i$ the variance of the Gaussian functions.

The RBF can be seen as a set of Gaussian functions which, through a weighting process, gives an evaluation of the class to which the instance $x$ belongs.

The connection with a special case of SVM methods can be done easily. In Eq. 6, if the type of the employed kernel is Gaussian, then the equivalence between SVM with Gaussian kernel and RBF is evident. In the case of the SVM with Gaussian kernel, the SVs represent in the original space, centres of Gaussian distributions. So the output of the method consists in a linear combination of Gaussian functions as in the case of RBF. The problem is to determine first the Gaussian components and second, the corresponding weights. As shown in [6], the centres of the Gaussian functions determined by SVM and by RBF formalisms correspond.

## BAYESIAN APPROACH

This section describes the Bayesian approach and the connections which can be established with SVM and Neural Network methods.

## Bayesian formalism

The Bayesian formalism is suitable for modelling high complexity data. The problems that usually arise when interpreting the data are to choose the correct model and to determine the right parameters of the model. Bayesian framework fits very well for accomplishing these two tasks by means of two level of inference.

The first level is assuming that the model which best suits the observed data $D$ is known. The problem is to find the set of parameters $w$ which corresponds to the considered model $M$. Using Bayes rule one can write:

$$p(w|D,M) = \frac{p(D|w,M)p(w|M)}{p(D|M)} \tag{8}$$

Infering the model's parameters of the data implies making assumptions on:

- the likelihood function $p(D|w,M)$ - how the data are generated for the assumed model
- the *a priori* information $p(w|M)$ - representing the prior belief of how parameters are best representing the correct model before any observation is done.

At the second level which is the model selection level, the Bayes rule is used to infer the model which best suits the observed data. The equation describing this level is given by the posteriori belief :

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)} \tag{9}$$

where the quantity $p(D|M)$, called the model likelihood or the evidence, is calculated by integrating over the space of model's parameters $w$ as $p(D|M) = \int p(D|w,M)p(w|M)dw$. The good model is the one with the highest posterior belief value. Considering the real case data with noise and the fact that the noise is normal distributed $\propto N(0,\sigma^2)$(as it will be considered for the rest of this paper) the considered model will map the instance $x$ into an output $y$ with the probability given by $p(y|w,x,M) = (\frac{1}{2\pi\sigma^2})^{1/2} exp(-\frac{1}{2\sigma^2}(y - g(x,w))^2)$.

The likelihood expression is obtained by $p(D|w,M) = \prod_{i=1}^{n} p(y_i|w,x_i,M)$. Taking into account the Bayes rule, the posterior distribution is proportional to:

$$p(w|D,M) \propto p(D|w,M)p(w|M) \tag{10}$$

An estimation of $w$ using the posterior distribution can be done by employing the Maximum a Posteriori (MAP) estimator. Maximising the expression in Eq. 10 is equivalent to minimising its negative logarithm, thus obtaining the following optimisation problem: $\min_w \quad \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}|y_i - g(x_i)|^2 + \Omega(w)\right)$ with $\Omega(w) = \frac{1}{2}||w||^2$ representing the regularization component.

## Bayesian versus SVM

A very nice connection can be established between SVM and Bayesian formalisms. This is due to the fact that probability measures can be attached to the SVs, thus allowing posterior probability measure as the output of the classification task [1, 13].

Moreover the classification task is done by solving a functional which is regularised. The choice of the regularization parameter and the kernel type can be done via the Bayesian perspective. Based on the infered models and parameters, the probabilistic class output can be generated.

In the case of probabilistic binary classification, the likelihood evaluation is given by:

$$p(y|g) = \frac{1}{1+exp(-y \cdot g)}. \tag{11}$$

The loss function defined as $l = -ln(p(y|g))$ indicates the loss in the classification process. In order to make the inference e.g. kernel type and parameters, a Bayesian framework is considered as described below. As before, $g$ is considered to be the result of random variables in a zero-mean Gaussian stochastic process. Thus it is described by the covariance matrix $\Sigma$. As presented in [1], the infered parameters are collected in a vector $w$ which gives the prior probability as in:

$$p(g|w) = \frac{1}{Z_g} exp(-\frac{1}{2} g^T \Sigma g) \tag{12}$$

with $g = [g(x_1), g(x_2), \cdots, g(x_n)]$, the covariance matrix $\Sigma$ and $Z_g = (2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}$. Introducing the loss function in the likelihood, one can obtain $p(D|g,w) = \prod_{i=1}^{n} p(y_i|g(x_i))$. Using the last two equations in Bayes formula, the posterior probability can be written as follows $p(g|D,w) \propto exp(\frac{1}{2} g^T \Sigma g + \sum_{i=1}^{n} l(y_i \cdot g(x_i)))$. By maximising it, the MAP estimator is obtained. The maximisation problem is equivalent with minimising the exponential factor as given below:

$$\min_{g} \quad \Big(\frac{1}{2} g^T \Sigma g + \sum_{i=1}^{n} l(y_i \cdot g(x_i))\Big) \tag{13}$$

It can be seen that the optimisation problem in Eq. 13 is similar with the one presented in the case of SVM expressed by Eq. 3. In a similar manner, Lagrange multipliers and slack variables are introduced and the dual problem is solved.

In order to infer $w$, the posterior probability $p(D|w)$is maximised. As this expression is not known, using Bayes rule one can solve the problem by maximising the likelihood function $p(D|w) = \frac{1}{Z_g} \int exp(-S(g))dg$ with $S(g) = \frac{1}{2} g^T \Sigma^{-1} g + \sum_{i=1}^{n} l(y_i \cdot g(x_i))$.

As mentioned before, only the SVs will be used in the estimation of $w$ instead of all $g_i$ determined coefficients. The classification can be done via probabilistic class prediction by computing $p(y|D,w)$ as presented in [1].

One important difference is that while the Bayesian is using all the training data to infer the model, the SVM is using only the determined SVs for the same purpose.

## Bayesian versus RBF

RBF represents a special case of Neural Networks with the decision function having the expression $f = sgn(g(x))$ with $g(x) = \sum_{i=1}^{n} w_i \cdot exp(-\frac{||x-x_i||^2}{c_i}) + b$.

In a similar manner, taking into account the connections between SVM and RBF, and the Bayesian representation of the SVM formalism, one can obtain a description of the RBF in Bayesian terms. It is possible to express the posterior distribution under the assumption of Gaussian noise in the same way as in the case of SVM:

$$log \quad p(w|D) = -\sum_{i=1}^{n}(y_i - g(x_i)) - \frac{1}{2}w^T w \tag{14}$$

Training RBF from Bayesian point of view is equivalent to the problem of infering the parameters involved in *g(x)*. However there are cases in RBF applications where the Gaussian parameters $c_i$ are considered constants and thus there is no need to infer them.

## MAXIMUM ENTROPY

Maximum entropy can be seen as a special case of Bayesian formalism and as well can be derived from SVM by introducing a special case of kernel function.

### MaxEnt versus Bayesian

The MaxEnt formalism introduced in [8, 12] and [10] is a method used to infer an unknown probability density function subject to a set of constraints. No *a priori* knowledge of the density functions is made. As previous, by denoting the data by vectors $x_i$, finding the probability density function $q^*$ which best describes the data with the imposed constraints, comes to finding among the constraints complying density function, the one with the highest entropy:

$$H(q) = -\sum_{i=1}^{n} q(x_i) log(q(x_i)) \tag{15}$$

The constraints imposed to the unknown probability density function are given as a set of expectations. The number of the imposed constraints is denoted by $m$:

$$\int \beta_k q^*(x)dx = \beta_k^* \tag{16}$$

with $k = 1 \cdots m$, $\beta_k$ and $\beta_k^*$ represent a set of known functions and a set of known constants respectively. The normalising condition is given by $\int q(x)dx = 1$.

Now considering the Lagrangian multipliers determined by the constraint equations, the solution obtained is given by $q(x) = p(x)Z^{-1}exp^{-\alpha\beta}$ with and $Z = exp(\mu)$ and $\mu$ given by $\mu = log(\int p(x)exp(-\sum_{k=1}^{m}\alpha_k\beta_k))dx$.

This is similar to the solution obtained in the case of SVM with a modified kernel function. Using the entropy concentration theorem [11], it can be checked that the possible distributions are concentrated strongly near the maximum value of entropy. Considering a random experiment with $N$ trials and each i-th result occurring $N_i = N \cdot \varphi_i$ times, $1 \leq i \leq n$, considering all possible $n^N$ outputs, the number which yields a particular set of frequencies $\varphi_i$ called the multiplicity factor is given by $W(\varphi_1, \cdots, \varphi_n) = \frac{N!}{(N\varphi_1!)\cdots(N\varphi_n!)}$. Using the Stirling approximation in the case of $N \to \infty$, it is obtained:

$$N^{-1}log(W) \to H \tag{17}$$

Considering two sets of frequencies $\varphi_i$ and $\varphi_i^{'}$ one obtains the qualitative expression of the entropy concentration theorem

$$\frac{W}{W'} \sim A \cdot exp(N(H - H^{'})) \tag{18}$$

It will be shown that the MaxEnt estimator is a particular case of the Bayesian inference case. One important aspect of the MaxEnt formalism is that it is not taking into account the noise present in the data. This is different from the Bayesian approach where the noise is taken into account and moreover its distribution is known and considered Gaussian. In order to incorporate the noise into MaxEnt formalism, the expression of constraints given in Eq. 16 is changed, by introducing the error vector $e_k = \int \beta_k(x) q^*(x) dx - \beta_k^*$ with $k = 1, \cdots m$. Considering the noise as having a normal distribution, $e_k \sim N(0, \sigma_k)$, the following quadratic form is defined $Q = \frac{1}{2} \sum_{k=1}^{m} \sigma_k^{-2} (\int \beta_k(x) q^*(x) dx - \beta_k)^2$. Taking into account the *a priori* information $I$ and the data $D$, the posterior probability of entropy $H$, employing Bayesian formalism is proportional to:

$$p(solution|D,I) \propto exp(NH - Q) \tag{19}$$

In Eq. 18 the prior probability is represented by the term *exp(NH)* while the likelihood by *exp(-Q)*. In the situation when the considered noise is absent (*Q=0*) the solution is similar with the one given by the MaxEnt formalism in Eq. 17. So it is shown that the MaxEnt formalism is included in the Bayesian one as a particular case. In the same way, the Bayesian results given by Eq. 19 can be interpreted as MaxEnt formalism. It can be pointed out that in a variational problem, a new constraint does not change the MaxEnt solution if it is already complying with the MaxEnt constraint. Equation 19 finds a maximum entropy $H$ for which the noise is at a level $Q_0$. This can be regarded as a maximisation problem of entropy $H$ with the constraint that the noise is maintained at the same level $Q_0$. Thus Bayesian formalism can be seen as a MaxEnt with constraints concerning the noise component.

## MaxEnt versus SVM

Another interesting link can be established between the SVM and MaxEnt formalisms.

Instead of using classical kernel function as in SVM, one can construct a new mapping procedure where the computation of the kernel function is done by employing a distance measure in the space of probability density functions [5]. This means that for each instance the probability density function is computed and then used in the classification process. As in the previous section, the transition from the classical kernel functions to the one which is based on probability density functions is described by $K(x_i, x_j) \rightarrow K(p(x|w_i), p(x|w_j))$, with $w_i$ representing as before the model's parameters and $p(x|w)$ can be assumed to be single full covariance Gaussian model as shown in [5].

As the feature space where the kernel function is computed is a statistical one, one can compute a distance: the Kullback-Leibler (KL) symmetric divergence in order to compare the two distributions. The expression is given by $KL(p(x|w_i), p(x|w_j)) = \int_{-\infty}^{\infty} p(x|w_i) log(\frac{p(x|w_i)}{p(x|w_j)}) dx + \int_{-\infty}^{\infty} p(x|w_j) log(\frac{p(x|w_j)}{p(x|w_i)}) dx$. Now in the expression of the Gaussian kernel, if the Euclidean distance is replaced with the new statistical one, one obtains $K(x_i, x_j) = exp(-\frac{KL(p(x|w_i), p(x|w_j))}{2\gamma^2})$. The obtained Kernel is a valid one [5] because the kernel matrix is a positive definite matrix and thus complies with the Mercer condition for kernels. Taking into consideration the optimisation problem presented in the case of SVM with Gaussian kernel by replacing the kernel with the new one, one

obtains an optimisation problem similar to the one presented in the case of MaxEnt formalism.

## DISCUSSIONS

The presented formalisms have a wide range of applications in image understanding systems. They are employed for classification, feature selection purposes and for Relevance Feedback (RF) problem in Content Based Image Retrieval System (CBIR).

SVM methods have been used in different RF systems with good results in discriminating relevant images within the database [7, 14]. In the same time, Bayesian RF algorithms have been proposed in [15]. RF algorithms are employed in CBIR retrieval systems in order to enhance the retrieval capabilities by human - machine interaction. The user is involved in the retrieval process by annotating the retrieved documents as relevant or irrelevant in order to increase the retrieval precision.

For comparison purposes, two data sets are considered as described below. The first one is obtained from gray level data of a SPOT5 scene. Each pixel in the image is described by three features corresponding to the three bands used (Red, Green and near Infrared). Overall, a total number of 200 points are used for representing two classes: water and forest. The second set corresponds to cropped SPOT5 scenes into small images with a dimension of 64x64 pixels. A total number of 200 small images representing classes of sea and city is obtained, with 100 images per class. The texture features of Quadrature Mirror Filters (QMF) (8) and Haralick matrix of co-occurrence (78) are used to describe the information within an image thus giving a 86-feature vector.

For each case, the data set was divided into two parts: one with 40 examples (20 examples per class) used for training procedure and the other with the rest of data used for retrieval purposes. Precision-Recall curves are considered as a good tool to evaluate the properties of a retrieval system. Let us denote the retrieved images by $A$ and the relevant ones by $B$. The precision $P$ is defined as the fraction of retrieved images which are relevant and the recall $R$ as the fraction of relevant images which have been retrieved: $P = \frac{|A \bigcap B|}{|A|}$ and $R = \frac{|A \bigcap B|}{|B|}$.

Figure 1 illustrates the Precision - Recall curves in the case of the previous described data when SVM based RF and Bayesian based RF are employed. On the left hand side of the picture, the curves are traced when using the first data set; and on the right hand side similar curves are ploted for the second data set.

From these results, it can be seen that the dimensionality of the feature vector does not influence the retrieval capabilities when employing SVM based RF. On the other hand, Bayesian based RF is affected by the high demensionality and in the case of 86-feature vector the retrieval capabilities are not so performant.

One major disadvantage of Bayesian methods over SVM methods used in RF is that for high dimensional data, Bayesian approach is not performing well, while SVM provides good results for highly dimensional data.

## ACKNOWLEDGMENT

The work was performed within The CNES/DLR/ENST Competence Centre on Information Extraction and Image Understanding for Earth Observation. SPOT5 images have been provided by "Centre National d'Etudes Spatiales" (CNES).

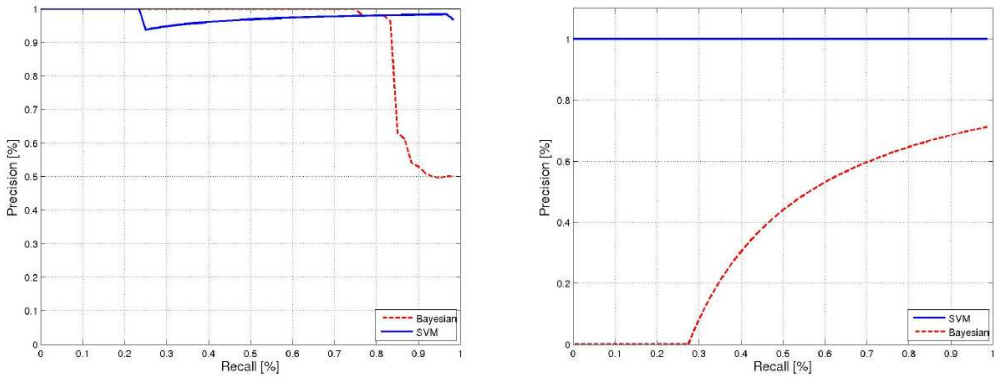


**Figure 1.** Precision - Recall curves for the case of 3 dimensional feature space (left hand side) and the case of 86 texture features (right hand side). The Bayes discriminant performs better for small dimensional data while SVM is not greatly influenced by the dimensionality of the data.

# REFERENCES


1. W. Chu and S. Sathiya Keerthi and C. J. Ong, A new bayesian design method for support vector classification, Proc. of IEEE Int'l Conf. on Multimedia and Expo, Lausane, Switzerland, 2002.
2. Christopher M. Bishop and Michael E. Tipping, Bayesian Regression and Classification, Advance in Learning Theory:Methods, Models and Applications, J.A.K. Suykens et al., IOS Press, NATO Science Series III: Computer and Systems Sciences, 190.
3. Christopher M. Bishop, Neural networks for pattern recognition, Oxford University Press., 1995.
4. K.-Robert Muller and S. Mika and G. Ratsch and K. Tsuda and B. Schöklkopf, An Introduction to Kernel-Based Learning Algorithms, IEEE Transactionson Neural Networks, Vol. 12, No. 12, March,2001
5. P. J. Moreno and P. P. Ho and N. Vasconcelos, A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications, Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, 2004.
6. B. Schöklkopf and Kah-Kay Sung and C. J. C. Burges and F. Girosi and P. Niyigi and T. Poggio and V. Vapnik, Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers, IEEE Transactions on Signal Processing, Vol. 45, No. 11, November, 1997.
7. M. Costache and H. Maitre and M. Datcu, Categorization based Relevance Feedback Search Engine for Earth Observation Images Repositories, IEEE International Geoscience and Remote Sensing Symposium, 2006.
8. S.F. Burch, S.F. Gull, J. Skilling, "Image Restoration by a Powerful Maximum Entropy Method", in *Comp. Vis. Graph. and Imag. Processing*, vol. 23, 113-128, 1983.
9. S.F. Gull, G.J. Daniell, The maximum entropy algorithm applied to image enhancement, in *IEE Proc.*, vol. 127E pp.170-172, 1980.
10. S.F. Gull, J. Skilling, Maximum Entropy Method in Image Processing, in *IEE Proc.*, vol. 131F, No.6, 646-659, 1984.
11. E.T.Jaynes, On The Rationale of Max-Ent Methods, *IEEE Proc.*, vol. 70, No.9, p. 939-952, 1982.
12. J.H. Justice (ed.), *Max. Entr. and Bayesian Methods in Appl. Statistics*, Cambridge U. Press, 1986.
13. MacKay, D. J. C, The evidence framework applied to classifictaion networks, Neural Computation, 4(5), 720-736, 1992.
14. M. Ferecatu and M. Crucianu and N. Boujemaa, Tuning SVM-Based relevance Feedback for the Interactive Classification of Images, Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT), 2004.
15. I. J. Cox and M. L. Miller and T. P. Minka and T. V. Papathomas and P. N. Yianilos, The Bayesian Image Retrieval System, PicHunter. Theory, Implementation and Psychophysical Experiments, IEEE Transactions on Image Processing, 2000, 20.