Overview of Bayesian Inference, Maximum Entropy and Support Vector Machines Methods

Mihai Costache[*], Marie Lienou[*] and Mihai Datcu[†,*]

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

Abstract. Discrimination, feature and model selection analyses from the perspective of classification performance conducts to the discussion of the relationships between Support Vector Machine (SVM), Bayesian and Maximum Entropy (MaxEnt) formalisms. Maximum Entropy 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 to a Maximum Entropy similarity.

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

INTRODUCTION

Discrimination, feature and model selection analyses from the perspective of classification performance conduct to the discussion of the relationships between different classification formalisms used in data mining, such as Support Vector Machine (SVM), Bayesian and Maximum Entropy (MaxEnt).

Therefore, each of the methods can be linked with the others in a particular manner. Thus each pair of formalisms is characterised by similarities and differences.

The present article illustrates these connections with an incursion in the history of the classification formalism, taking into consideration the evolution from the original simple linear classifier, e.g. 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 differences concerning the two steps. In order to keep it as clear and simple as possible we will define our learning problem in the case of the binary classification. Thus the task is to find the decision function f which, based on independent observations, assigns an instance x to one of the two classes denoted by {+1,-1}.

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

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

where '·' 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 RBF formalisms, then 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 can be derived from SVM by means of Kernel function. In the last section, discussions concerning the practical aspects of SVM and Bayesian methods are presented.

SUPPORT VECTOR MACHINES

The SVM principle has his roots in the well known Perceptron formalism.

The Perceptron is the first binary linear classifier and represents the simplest kind of feedfoward Neural Network. 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 weights vector w and the scalar b starting from a training sequence.

Different algorithms are employed for this purpose: Stochastic Gradient Descent, Mean Square Error and Cross-Entropy.

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 to trace two surfaces that best delimitate the examples in the 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 $f(x) = sgn(w \cdot \Phi(x) + b)$, we 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, the goal is to minimise the expected risk as stated in [3], thus $||w||^2$. The problem is translated into a margin area maximisation, as the margin area is given by $\frac{2}{||w||}$. Introducing the Lagrange multipliers α_i, $i = 1, \cdots, n$ for each of the conditions in Eq. 2 we obtain:

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

Minimising Eq. 3 in respect to w and b and maximising it in 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 we obtain the dual quadratic optimisation problem:

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

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, the Lagrange multipliers $\alpha_i, i = 1, \cdots, n$ are obtained and used to compute the decision function used for classification as follows:

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

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, we 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. They contain class examples in order to obtain the decision function.

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 Radial Basis Function (RBF) is considered in this paper.

Radial Basis Function

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

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

with x_i representing the centre and c_i the variance of 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 [5] 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 modeling complex 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 assuming that the model which best suits the observed data, D, is known. The problem is to find the set of parameters θ which corresponds to the considered model M. Using Bayes rule we have:

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

Infering the model's parameters from data implies making two assumptions:

- likelihood function $p(D|\theta,M)$ - how the data are generated for the assumed model
- prior information $p(\theta|M)$ - 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 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 θ as $p(D|M) = \int p(D|\theta,M)p(\theta|M)d\theta$. 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 $p(D|w,M) = \prod_{i=1}^{n} p(y_i|w,x_i,M)$ where the parameters vector θ is replaced with w. Taking into account the Bayes rule, we derive the posterior distribution as being 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 $\min_w \left(\frac{1}{2\sigma^2} \sum_{i=1}^{n} |y_i - g(x_i)|^2 + \Omega(w) \right)$ with $\Omega(w)$ 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. 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 Bayesian perspective. Based on the infered models and parameters, 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 $l = -ln(p(y|f))$ 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. The functional 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 Σ. As presented in [1] the infered parameters are collected in a vector θ which gives the prior probability as in:

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

with $g = [g(x_1), g(x_2), \cdots, g(x_n)]$, the covariance matrix Σ and $Z_g = (2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}$.

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

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

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 θ, the posterior probability $p(D|\theta)$ is maximised. As we do not have this expression, using Bayes rule we can solve the problem by maximising the likelihood function $p(D|\theta) = \frac{1}{Z_g} \int exp(-S(g)) dg$ with $S(g) = \frac{1}{2} g^T \Sigma^{-1} g + \sum_{i=1}^{n} l_t(y_i \cdot g(x_i))$.

As mentionned before, only the SVs will be used in the estimation of θ instead of all g_i determined coefficients. The classification can be done via probabilistic class prediction by computing $p(y|D, \theta)$ 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, we can obtain a description of the RBF in Bayesian terms. We can express the posterior distribution under the assumption of Gaussian noise in the same way as in the case of SVM:

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

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 shape of the Gaussian components 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 - Bayesian

The MaxEnt formalism introduced in [7], [11] and [9] 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:

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

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:

$$\int \beta_k q^*(x)dx = \beta_k^* \qquad (16)$$

with $k = 1 \cdots m$, β_k and β_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 $q(x) = p(x)Z^{-1}exp^{-\alpha f}$ with and $Z = exp(\mu)$ and μ given by $\mu = log(\int p(x)exp(-\sum_{k=1}^{m}\alpha_k f_k))dx$.

This is similar to the solution obtained in the case of SVM with changed Kernel function. Using the entropy concentration theorem [10] 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 $N_i = N \cdot f_i$ times, $1 \le i \le n$, considering all possible n^N outputs, the number which yields a particular set of frequencies f_i called the multiplicity factor is given by $W(f_1, \cdots, f_n) = \frac{N!}{(Nf_1!)\cdots(Nf_n!)}$. Using the Stirling approximation in the case of $N \to \infty$, we get:

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

Considering two sets of frequencies β_i and β_i' we obtain the qualitative expression of the entropy concentration theorem

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

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, we must change the expression of constraints given in Eq. 16 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 $Q = \frac{1}{2}\sum_{k=1}^{m} \sigma_k^{-2}(\int \beta_k(x)q^*(x)dx - \beta_k)^2$. Considering the prior 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 we have 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 final solution if it is already complying with the 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. So Bayesian formalism can be seen as a MaxEnt with constraints concerning the noise component.

MaxEnt - SVM

Another interesting link can be done between the SVM and MaxEnt formalisms. Instead of using classical kernel function as in SVM, we 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 function (pdf) functions [4]. This means that for each instance the pdf is computed and then used in the classification process. As in the previous section the transition from the classical kernel functions to one which is based on pdfs is described by $K(x_i, x_j) \rightarrow K(p(x|\theta_i), p(x|\theta_j))$ with θ_i representing as before the model's parameters and $p(x|\theta)$ can be assumed to be single full covariance Gaussian model as shown in [4].

As the feature space where the kernel function is computed is a statistical one, we compute a distance: the Kullback-Leibler (KL) symmetric divergence in order to compare the two distributions. The expression is given in $KL(p(x|\theta_i), p(x|\theta_j)) = \int_{-\infty}^{\infty} p(x|\theta_i log(\frac{p(x|\theta_i)}{p(x|\theta_j)}))dx + \int_{-\infty}^{\infty} p(x|\theta_j log(\frac{p(x|\theta_j)}{p(x|\theta_i)}))dx$. Now in the expression of the Gaussian kernel, if the Euclidean distance is replaced with the new statistical one, we obtain $K(x_i, x_j) = exp(-\frac{KL(p(x|\theta_i), p(x|\theta_j))}{2\gamma^2})$. The obtained Kernel is a valid one [4] 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, we obtain 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 [6]. In the same time, Bayesian RF algorithms have been proposed in [12]. RF algorithms were employed first in text retrieval systems in order to enhance the retrieval capabilities by human - machine interactions. 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 the next step.

The SVM based RF algorithms tries to find a separation surface in the feature space which best delimitates the relevant / irrelevant examples and thus using only examples which are near the already traced surface. The Bayesian approach is using examples spread all over feature space in order to extract parameters of models. Once these models are constructed, they are used to discriminate the annotated examples.

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 great results for highly dimensional data.

For this purpose RF process is simulated in two cases: the first consisting in data set obtained from gray level data of SPOT images used to generate Corine Land Cover (CLC) maps. Each pixel in the image is described by three features corresponding to the three bands used. Overall, a total number of 200 points are used for representing two CLC clases: water and forests. The data set was divided into two parts: one with 40 examples for each class used for training procedure and the second one with the rest of data used for retrieval purposes.

In the second case SPOT5 images with a dimension 64×64 representing classes of sea and city are used. The texture features of Quadrature Mirror Filters (QMF) and Haralick matrix of co-occurrence are used, obtaining a feature vector of 86 features. 100 examples per class are employed with 20 examples used for learning step and the rest of 80 for retireval purposes. In a similar manner, RF process is simulated. 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 the retrieved images which are relevant and the recall R as the fraction of relevant images which have been retrieved: $P = \frac{|A \cap B|}{|A|}$ and $R = \frac{|A \cap B|}{|B|}$.

Figures 1 illustrates the retrieval properties in the case when either Bayesian or SVM methods are used. The Bayesian approach used to retrieve the documents is performing better for smaller Recall values than the case where SVM is used. It can be seen that in the case of SVM, the retrieval properties are good so the system is capable of well identifying and retrieving the relevant images, while in the case of Bayesian retrieval, the performances are quite bad.
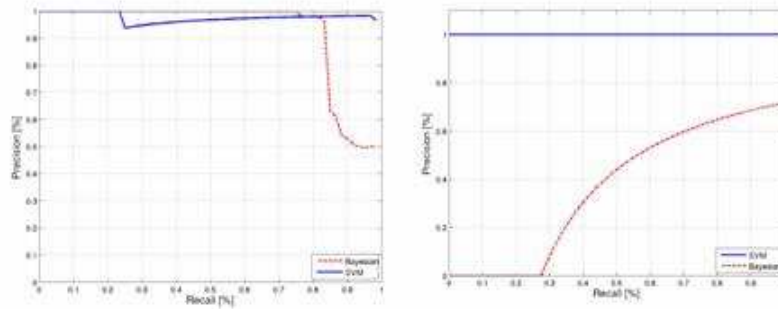
ACKNOWLEDGMENT

FIGURE 1. Precision - Recall curves for the case of 3 dimensional feature space (left hand side) or the case of 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.

been provided by "Centre National d'Etudes Spatiales" (CNES).

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. 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
4. 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.
5. 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.
6. Categorization based Relevance Feedback Search Engine for Earth Observation Images Repositories, M. Costache and H. Maitre and M. Datcu, IEEE International Geoscience and Remote Sensing Symposium, 2006.
7. 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.
8. S.F. Gull, G.J. Daniell, "The maximum entropy algorithm applied to image enhancement", in *IEE Proc.*, vol. 127E pp.170-172, 1980.
9. S.F. Gull, J. Skilling, "Maximum Entropy Method in Image Processing", in *IEE Proc.*, vol. 131F, No.6, pp.646 - 659, 1984.
10. E.T.Jaynes, "On The Rationale of Max-Ent Methods", *IEEE Proc.*, vol. 70, No.9, p. 939-952, 1982.
11. J.H. Justice (ed.), *Max. Entr. and Bayesian Methods in Appl. Statistics*, Cambridge U. Press, 1986.
12. 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.