Clustering with Normalized Information Potential Constrained Maximum Entropy Boltzmann Distribution

Umut Ozertem, Deniz Erdogmus

Abstract — **From a probabilistic perspective, the question of clustering is '*what is the probability that two data samples belong to the same cluster?*' Accepting the natural preclustering of samples into corresponding modes of the data probability distribution and answering the question posed above for these modes can reduce the problem complexity. Under the maximum entropy principle, a Boltzmann distribution model can be employed to evaluate the required mode-connectivity probabilities. An algorithm is developed using kernel density estimation in this framework. Its performance is demonstrated on benchmark datasets.**

I. Introduction

CLUSTERING is a fundamental problem in machine learning that has a wide range of applications in data analysis and mining, as well as compression and noise reduction [1,2,3,4]. Literature is rich with many algorithms, heuristic and principled; hierarchical clustering and k-means are arguably the most popular methods due to their simplicity and have found widespread use Nevertheless, the former might suffer from high statistical variance if samples are slightly perturbed and the latter is motivated by the Gaussian assumption of clusters and minimum information loss in compression. More recently, affinity-based clustering algorithms other than hierarchical clustering have gained popularity, exactly due to the possible problem mentioned above. For example, the normalized cut algorithm [5] measures the similarity between two partitions using a normalized average similarity measure much like the correlation coefficient. In fact, it has been shown that the normalized cut measure and related spectral clustering algorithms are attempting to maximize the Cauchy-Schwartz (CS) distance (angular separation) between the partitions [5]. This distance is essentially a normalized information potential measure in the context of nonparametric entropy estimation as we will discuss later in the paper. There have also been perturbation approaches [6] to counter the problem mentioned above. The idea is to create a mask for the affinity matrix based on the connectivity probabilities of pairs of samples in proximity graphs (such as k-nearest-neighbor or ε-ball graphs) when the samples are *slightly* perturbed. This perturbation approach can in fact be interpreted as convolving the empirical data distribution locally with the perturbation distribution.

The probabilistic approach to clustering continuous-valued feature vectors demands that the inference is made in accordance with the underlying probability density function (pdf) of the data. In practice, the true pdf of the data is rarely known, so it must be estimated using a suitable approximation technique. In many complex domains parametric modeling may not be a feasible option due to model complexity selection issues. Nonparametric density estimation provides an attractive alternative; especially kernel density estimation (KDE) offers an appealing connection with the spectral clustering methodology.

Mean-shift procedure provides a means for achieving density-based cluster inference [7,8,9]; all samples in the attraction basin of a mode of the kernel density estimate (also KDE) are assigned to one cluster. This method has the drawback of potentially resulting in many small clusters caused by "statistically insignificant" modes – a situation that occurs more severely if the *optimal* kernel size is underestimated. Comaniciu and colleagues [9] proposed a heuristic approach to merge modes resulting from the mean-shift procedure by considering the distance between the peaks of the modes. We recently proposed merging nearby modes using the CS distance between pairs of modes [10], which expectedly results in more robust merging decisions taking the amount of spread and overlap of the clusters into account. Still, one needs to decide on a specific threshold value to draw the line between close and far modes given the CS distances.

In this paper, inspired by the typical cut algorithm [11,12], we develop a Boltzmann distribution prior for the mode connectivity probabilities. The Boltzmann distribution is the maximum entropy distribution under the constant average energy constraint, where energy in this case is defined to be the normalized information potential.

II. Data Model For Clustering

The modes of the probability distribution underlying the data provide a natural clustering structure that is in accordance with the probabilistic view of data. It is important, however, to note that each mode may not be a real or statistically significant cluster, especially when the distribution is estimated from samples. This is most likely the case in practice, since true underlying data distributions are generally unknown. Kernel density estimation provides a

The authors are with the Computer Science and Electrical Engineering Department, Oregon Health and Science University, Portland, OR 97239 USA, (e-mails: [deniz, ozertemu]@csee.ogi.edu).

convenient technique for nonparametrically estimating the probability distribution and finding the *natural* clusters.

A. Kernel Density Estimation

Given the sample set $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, the KDE using the kernel function $K(\boldsymbol{\xi})$ is given by [13,14]

$$p(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} K_{\boldsymbol{\Sigma}_i}(\mathbf{x}-\mathbf{x}_i) \tag{1}$$

where the kernel function is typically a valid pdf itself to guarantee the validity of the estimate in (1). We further impose the continuity and differentiability conditions on the kernel for an iterative gradient-based mode finding algorithm to be possible. Typical kernels used are uniform and Gaussian distributions. The subscript of the kernel function denotes the parameter that controls its size also known as the bandwidth).

There are mainly two options for kernel size selection: fixed and variable. Fixed kernel size, $\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}$, has the convenience of reduced model complexity, however it is also problematic and insufficient when the data exhibits distributions with multiple scales. Combination kernels are recently proposed as a viable solution for this problem [15]. Variable-size kernels are more flexible in their capability to model various scales present in the data and certain simplifying assumptions could be made. For example, the kernel size associated with a specific sample could be limited to a scaled version of the k-nearest neighbor (KNN) covariance, that is $\boldsymbol{\Sigma}_i=\sigma^2\boldsymbol{C}_i$. Both in the case of fixed and constrained variable kernel size cases, the global scale parameter σ can be determined to optimize a suitable selection criterion. For example, Silverman's rule [16] could be used to select a kernel size that minimizes the squared error while maintaining the average curvature of the estimated density equal to that of a Gaussian, or the leave-one-out maximum likelihood approach [17] could be employed to find a suitable kernel size. Kernel selection is an important consideration for all kernel-based solutions and the connection to kernel density estimation provides a principle for tackling this question.

B. Fixed Point Algorithm for Mode Finding

The modes of the pdf estimate in (1) can be easily obtained by an iterative mode search-algorithm. Due to the continuity and differentiability assumptions, the modes satisfy $\nabla_{\mathbf{x}} p(\mathbf{x}) = \mathbf{0}$. A fixed-point iteration can be easily obtained for any suitable kernel. Specifically for Gaussian kernels, the following iteration is obtained from (1):

$$\mathbf{x} \leftarrow \frac{\sum_{i=1}^{N} \mathbf{x}_i G_{\boldsymbol{\Sigma}_i}(\mathbf{x}-\mathbf{x}_i)}{\sum_{i=1}^{N} G_{\boldsymbol{\Sigma}_i}(\mathbf{x}-\mathbf{x}_i)} \tag{2}$$

It can be shown that this procedure, called mean-shift [18,19], can also be viewed as an expectation-maximization (EM) procedure to find the translation of $p(\mathbf{x})$ that maximizes the likelihood of the initial condition [19]. Consequently, this iteration converges linearly to the local maximum whose attraction basin contains the initial condition.

In clustering, we are interested in finding which mode a particular sample belongs to. Therefore, the mean shift procedure in (2) must be performed initializing it to all samples in the data set. Note that, this process will not necessarily find all the modes of the pdf, as in high dimensional mixture models such as (1), modes can appear at locations other than the component centers [19].

Recall that not all the modes of the estimated probability distribution are significant clusters; this is especially true if the kernel size is selected to be small. Alternatively, there might exist modes that are connected to each other with a significantly large probability density bridging them. One needs a mechanism to determine whether some neighboring modes determined using (2) should be merged into a single cluster or not. We propose to make this decision in a manner similar to the typical cut formalism, using a maximum entropy distribution model for the mode-connectivity. Under a constant average energy constraint, this results in a Boltzmann distribution over which we can evaluate the likelihood of two modes being in the same cluster.

C. Maximum Entropy Principle and the Boltzmann Distribution

In statistical mechanics, the Boltzmann distribution emerges as the solution to a constrained maximum entropy problem. Accoding to the maximum entropy principle [20], the distribution that best models the data should match observed properties exactly, while allowing maximum uncertainty.

Fact: Let the energies for all possible states $\{s_1,\ldots,s_n\}$ that the particle system can take be $\{E_1,\ldots,E_n\}$. Then, the probability distribution of the states that have maximum entropy under the constant average energy constraint is the Boltzmann distribution: $p(s_i)=exp(-E_i/T)/Z_T$.

Proof: The problem is formally stated as

$$\max_p -\sum_i p_i \log p_i \;\text{ s.t. }\; \sum_i p_i = 1 \;\text{ and }\; \sum_i E_i p_i = \overline{E} \tag{3}$$

The Lagrangian for this problem is

$$\Im = -\sum_i p_i \log p_i - \lambda_0\left(\sum_i p_i - 1\right) - \lambda_1\left(\sum_i E_i p_i - \overline{E}\right) \tag{4}$$

Derivating (4) with respect to p_i and equating to zero yields $\log p_i = 1+\lambda_0+\lambda_1 E_i$. The Boltzmann distribution is obtained by letting $T=-1/\lambda_1$ and $Z_T^{-1}=\exp(1+\lambda_0)$. □

In the Boltzmann distribution, T denotes the *ambient temperature* and can be adjusted to yield various distributions. Mathematically, changing the temperature corresponds to changing the average energy value in the constraint, as also expected intuitively.

D. Normalized Information Potential

The mean shift procedure described above provides a preliminary vector quantization solution for clustering. A

robust measure of similarity between two modes is given by the CS inequality in the pdf space. Using the standard integral-based inner product definition, an angular measure of distance between two pdfs $p_i(\mathbf{x})$ and $p_j(\mathbf{x})$ is given by:

$$\widetilde{K}(p_i, p_j) = \frac{< p_i, p_j >}{\| p_i \| \cdot \| p_j \|} \tag{5}$$

This measure is referred to as the CS-distance or normalized cross-information potential in previous work by Principe and colleagues and has been used successfully in blind source separation, dimensionality reduction, and clustering [22].

Suppose that the samples $\{\mathbf{x}_1^i,...,\mathbf{x}_{N_i}^i\} \subset \{\mathbf{x}_1,...,\mathbf{x}_N\}$ are the samples that are associated with mode i, the inner product between the pdfs of two modes can be calculated as

$$\begin{aligned} < p_i, p_j > &= \int p_i(\mathbf{x}) p_j(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \int G_{\mathbf{\Sigma}_k^i}(\mathbf{x} - \mathbf{x}_k^i) G_{\mathbf{\Sigma}_l^j}(\mathbf{x} - \mathbf{x}_l^j) d\mathbf{x} \\ &= \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} G_{\mathbf{\Sigma}_k^i + \mathbf{\Sigma}_l^j}(\mathbf{x}_l^j - \mathbf{x}_k^i) \end{aligned} \tag{6}$$

For fixed-size kernels, the inner product in (6) is identical to the popular graph cut measure [21], hence the CS distance in (5) becomes equivalent to the normalized cut [5]. Normalizing the graph cut is known to lead to more balanced clustering solutions by penalizing small clusters that are weakly connected to the rest of the data. Consequently, the measure in (5) not only considers the connection strength but also the cluster volume, as the norm is referred to in the spectral clustering literature. In the context of information theory, the norms are related to Renyi's quadratic entropy (hence the name normalized cross information potential). Specifically, Renyi's quadratic entropy of a cluster is [22]:

$$H_2(p_i) = -\log \int p_i^2(\mathbf{x}) d\mathbf{x} = -\log \| p_i \|^2 \tag{7}$$

Larger values of the measure in (5) occur when two modes are strongly connected according to (6) and when both modes have large quadratic entropies (that is volume).

E. Assessing Mode Connectivity with Mean Normalized Information Potential

For the given data set $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, suppose that the mean shift procedure results in M modes represented by their peak locations $\{\mathbf{c}_1,\ldots,\mathbf{c}_M\}$. Let $\wp$ be the set of all possible m-way partitions (allowing empty partitions) of these modes; this set contains m^M elements denoted by $\mathbf{q}_s$, where $\mathbf{q}_s$ is an M-vector consisting of q_{sk} that are the labels 1 to m assigned to mode $\mathbf{c}_i$ in partition $\mathbf{q}_s$. In this notation, the *energy* of the particular partitioning scheme $\mathbf{q}_s$ is given by:

$$E_s = \sum_{k=1}^{M} \sum_{l=1}^{M} \widetilde{K}_{kl} (1 - \delta_{q_{sk} q_{sl}}) \tag{8}$$

Under the constant-average energy assumption, the maximum entropy probability distribution for all partitions in $\wp$ is given by the Boltzmann distribution: $exp(-E_s/T)/Z_T$.

The assessment of modes $\mathbf{c}_i$ and $\mathbf{c}_j$ being in the same cluster is based on the expected value of these modes being in the same partition according to the Boltzmann distribution:

$$p_{ij} = \sum_s \delta_{\mathbf{q}_{si} \mathbf{q}_{sj}} p_S = \frac{\sum_s \delta_{\mathbf{q}_{si} \mathbf{q}_{sj}} e^{-E_s/T}}{\sum_s e^{-E_s/T}} \tag{9}$$

For small number of modes (determined by the dataset and the selected kernel size) and small m (selected by the user), the expectation in (9) can be calculated analytically. The same is not true for the original dataset in general. In the typical cut algorithm, the authors employ Swendsen-Wang (SW) sampling to generate samples from the target Boltzmann distribution in order to obtain a sample-average estimate of (9). In the event of a large number of modes, the sampling approach could be employed as well. It is expected, however, that if the same number of samples are generated using the SW technique for the Boltzmann distribution defined over all possible partitions of sample connectivities and mode connectivities, the latter will yield significantly lower statistical variance since the support of the latter discrete distribution is much smaller than that of the former.

III. RESULTS

In this section, we will demonstrate results on benchmark datasets (handwritten digits and image segmentation) using the proposed method and provide comparisons with the typical cut algorithm.

A. Handwritten Digits Dataset

The data can be found in the UCI Data Repository [23]. For illustrative purposes here we use only two classes (zeros and ones) from this data set. In total we randomly select 640 samples, equally from each class. Starting from each sample, we perform mean shift iterations as in (2) using a variable-size KDE with $\mathbf{\Sigma}_i = \sigma^2 \mathbf{C}_i$ to exploit the locality information of the data. The KNN approach (with 10 neighbors) is used for kernel size selection and results are presented for various values of the global scale parameter.

Figure 1 shows the sample-affinity and expected mode-affinity matrices obtained with two values of the global scale parameter. The expected mode connectivity matrices are also mapped back to the sample indices to demonstrate th corresponding expected sample affinities resulting from the proposed procedure.

Figure 2 shows the variation in the number of clusters identified by the expected mode affinity method as the temperature is varied. It is desirable that the *correct* number of classes survives for a large interval of temperature values.

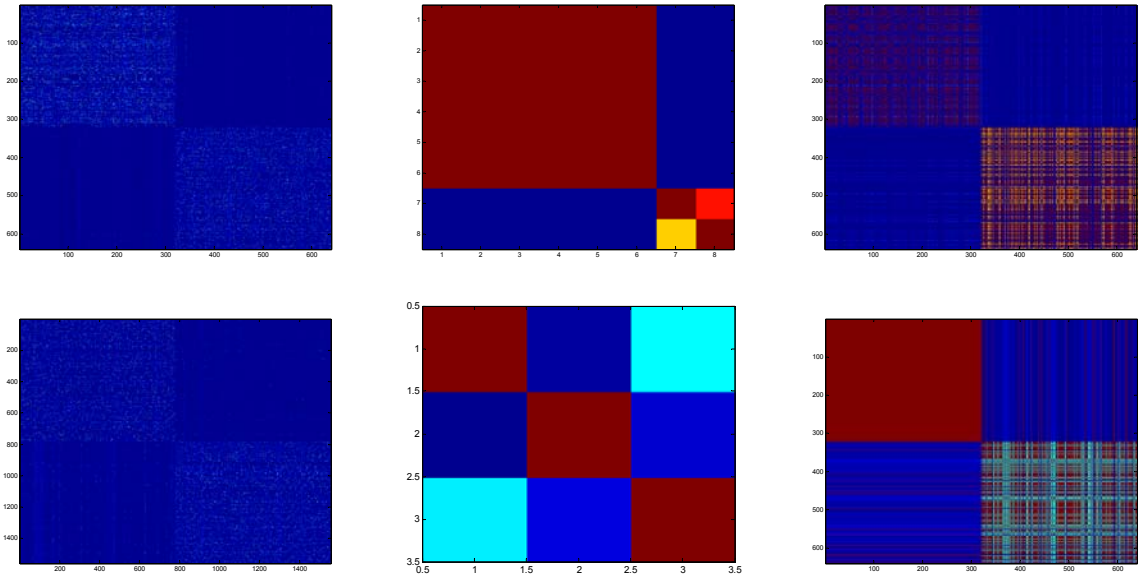

Figure 1. The sample affinity matrix (left), normalized mode affinity matrix (middle), and normalized mode affinity mapeed back to the sample pairs (right) are shown for σ = 0.5 (top) and σ = 1 (bottom).

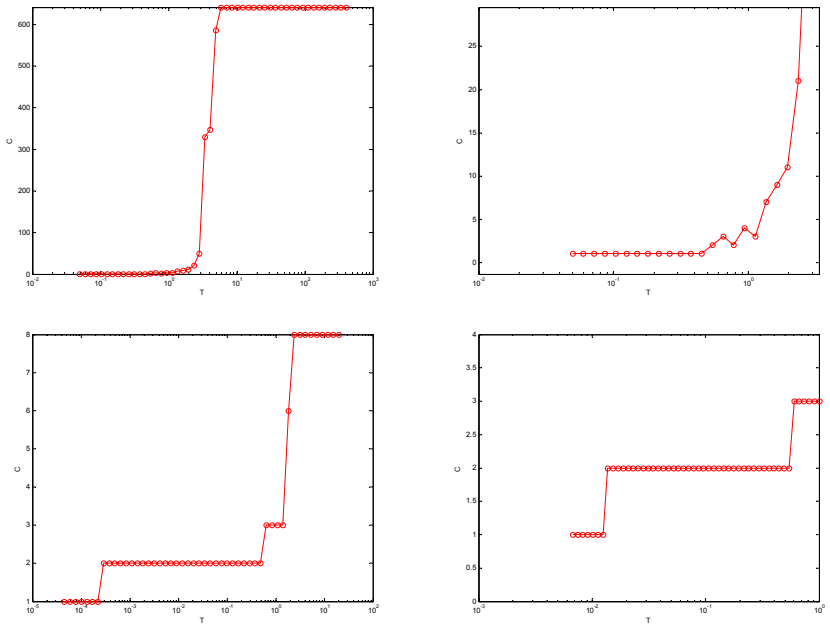

Figure 2. Number of clusters versus temperature (T) employing typical cut with global scale factor σ = 0.5, full range of T on the left and region of interest (around two clusters) on the right at the top row. The C vs T plots for mode-affinity approach σ = 0.5 (corresponds to Fig. 1 top-middle) and σ = 1 (corresponds to Fig. 1 bottom-middle) are shown at the bottom row.

TABLE 1 CONFUSION MATRICES FOR TWO GLOBAL SCALE VALUES

$$P(\sigma = 1) = \begin{bmatrix} 0.975 & 0.025 \\ 0.043 & 0.9563 \end{bmatrix} \qquad P(\sigma = 0.5) = \begin{bmatrix} 0.9594 & 0.0406 \\ 0.0625 & 0.9375 \end{bmatrix}$$

We observe that the sample-based typical cut procedure obtains the desired two-cluster solution briefly along the temperature axis, while the expected mode affinity method determines that thre are two clusters for a wide range of temperature values. The minimum number of clusters is one, obtained at the lower extreme of temperature, and the maximum number of clusters is the number of particles (modes or samples), obtained at the higher extreme of

(a) (b)

Figure 3. (a) The original image, (b) Segments obtained by mean shift (different grayscale levels) and the clustering boundary after merging according to the Boltzmann distribution (bright white curve).

temperature. Another important observation is that regardless of the global scale value assumed for the kernel size, the probability of correct classification of samples to the true classes remain unaffected. This is shown in the form of confusion matrices in Table 1.

B. Image Segmentation

Large number samples lead to affinity matrices whose size makes pair-wise affinity based clustering methods for image segmentation intractable, therefore typically approximate solutions based on the larger eigenvectors of the sample similarity matrix are employed [24]. Employing typical cut in such situations is next to impossible. Mean-shift provides a natural vector quantization solution as an alternative to pixel-wise computations, where the image is first transformed into a feature domain, which is usually obtained by using the x-y coordinate values and the corresponding intensity values I(x,y) for each channel of the image. Mean shift leads to many insignificant clusters that correspond to numerous modes in the KDE. The proposed method can be employed to merge such small segments. This is illustrated on the plane image that can be found in the Berkeley Image Segmentation Database [25]. Particularly, the image used in the experiments is 96×128 and results in a 12288×12288 pixel-affinity matrix. The original image is shown in Figure 3a and the results of mean-shift iterations in Figure 3b, along with the decision boundary for the two-cluster mode merging solution using the expected mode connectivity values over the Boltzmann distribution.

IV. CONCLUSIONS

In this paper, we approach the question of two points being in the cluster in a probabilistic perspective using the maximum entropy principle. We obtain significant computational savings by utilizing the mean-shift procedure as a vector quantization step and considering the modes of the density as a natural intermediate clustering solution. In practice, this also extends the application areas of typical cut algorithm into larger datasets. Moreover, the results show less dependency on perturbation on the temperature, which can be considered as introducing a sense of generalization to the data with the vector quantization. Eliminating the necessity for sampling for the pdf of possible states or reducing the statistical variance resulting from this sampling, utilization of the proposed quantized affinity matrix leads the statistically important clustering solutions to remain unchanged for a wider range of T and helps the corresponding clustering solution to be observed much more easily.

ACKNOWLEDGMENTS

This work was partially supported by NSF grant ECS-0524835. The authors would like to thank Miguel Carreira-Perpinan for valuable discussions.

REFERENCES

[1] P. Chang, D. Schlag, J. Zien, "Spectral K-Way Ratio-Cut Partitioning and Clustering," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 13, no. 9, pp. 1088-1096, 1994.

[2] Z. Wu, R. Leahy, "An Optimal Graph Theoretic Approach to Data Clustering: Theory and its Application to Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1101-1113, 1993.

[3] J. Goldberger, H. Greenspan, S. Gordon, "Unsupervised Image Clustering Using the Information Bottleneck Method," Proceedings of the DAGM Symposiumon Pattern Recognition, pp. 158-165, 2002.

[4] L. Hagen, A.B. Kahng, "Fast Spectral Methods for Ratio Cut Partitioning and Clustering," Proceedings of the International Conference on Computer-Aided Design, pp. 10-13, 1991.

[5] J. Shi, J. Malik, “Normalized Cuts and Image Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 888-905, 2000.
[6] M.A. Carreira-Perpinan, R.S. Zemel, “Proximity Graphs for Clustering and Manifold Learning,” Proceedings of NIPS’04, 2005.
[7] Y. Cheng, “Mean Shift, Mode Seeking, and Clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 8, pp. 790-799, 1995.
[8] D. Comaniciu, P. Meer, “Mean Shift: A Robust Approach Toward Feature Space Analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603-619, 2002.
[9] B. Georgescu, I. Shimshoni, P. Meer, “Mean Shift Based Clustering in High Dimensions: A Texture Classification Example,” Proceedings of the International Conference on Machine Vision, pp. 456-463, 2003.
[10] U. Ozertem, D. Erdogmus, "Spectral Clustering with Mean Shift Preprocessing," Proceedings of MLSP, pp. 73-78, Sep. 2005.
[11] M. Blatt, S. Wiseman, and E. Domany, “Data clustering using a model granular magnet,” Neural Computation, vol. 9, no. 8, pp. 1805.1842, Nov. 1997.
[12] Y. Gdalyahu, D. Weinshall, and M. Werman, “Self organization in vision: Stochastic clustering for image segmentation, perceptual grouping, and image database organization,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1053.1074, Oct. 2001.
[13] E. Parzen, “On Estimation of a Probability Density Function and Mode”, in Time Series Analysis Papers, Holden-Day, Inc., San Diego, California, 1967.
[14] L. Devroye, G. Lugosi, Combinatorial Methods in Density Estimation, Springer, New York, 2001.
[15] A. J. Smola, R. C. Williamson, and B. Schölkopf. “Generalization bounds for convex combinations of kernel functions,” NeuroCOLT Technical Report NC-TR-98-022, Royal Holloway College, University of London, UK, 1998.
[16] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.
[17] Duin R.P.W. 1976. On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Transactions in Computing C-25: 1175–1179.
[18] M. Á. Carreira-Perpiñán, “Mode-finding for mixtures of Gaussian distributions”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1318-1323, 2000.
[19] M. Á. Carreira-Perpiñán, and C. K. I. Williams, “On the number of modes of a Gaussian mixture”, Scale-Space Methods in Computer Vision, pp. 625-640, Lecture Notes in Computer Science vol. 2695, Springer-Verlag.
[20] Jaynes, E. T., “Where Do We Stand On Maximum Entropy” in The Maximum Entropy Formalism, R. D. Levine and M. Tribus Eds., pp. 15-118, Cambridge: MIT Press, 1978
[21] S. Sarkar, P. Soundararajan, “Supervised Learning of Large Perceptual Organization: Graph Spectral Partitioning and Learning Automata,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 52, pp. 504-525, 2000.
[22] D. Erdogmus, “Information Theoretic Learning Renyi’s Entropy and its Applications to Adaptive System Training,” PhD Dissertation, University of Florida, 2002.
[23] Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
[24] Y. Weiss, “Segmentation Using Eigenvectors: A Unifying View,” International Conference on Computer Vision, pages 975-982, 1999.
[25] D. Martin and C. Fowlkes and D. Tal and J. Malik, “A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics,” Proc. 8th Int'l Conf. Computer Vision, vol. 2, pp. 416-423, July 2001.