

INFORMATION CUT AND INFORMATION FORCES FOR CLUSTERING *

Robert Jenssen^{1,2†}, Jose C. Principe¹ and Torbjørn Eltoft²

¹ Computational NeuroEngineering Laboratory
University of Florida, Gainesville, FL. 32601, USA

² Electrical Engineering Group
University of Tromsø, 9037 Tromsø, Norway

Abstract. We define an information-theoretic divergence measure between probability density functions (pdf's) that has a deep connection to the cut in graph-theory. This connection is revealed when the pdf's are estimated by the Parzen method with a Gaussian kernel. We refer to our divergence measure as the Information Cut. The Information Cut provides us with a theoretically sound criterion for cluster evaluation. In this paper we show that it can be used to merge clusters. The initial clusters are obtained based on the related concept of Information Forces.

We create directed trees by selecting the predecessor of a node (pattern) according to the direction of the Information Force acting on the pattern. Each directed tree correspond to a cluster, hence enabling us to obtain an initial partitioning of the data set. Subsequently we utilize the Information Cut as a cluster evaluation function to merge clusters until the predefined number of clusters is reached.

We demonstrate the performance of our novel information-theoretic clustering method when applied to both artificially created data and real data, with encouraging results.

INTRODUCTION

Clustering is the problem of partitioning a set of data patterns into different subsets, such that patterns within each subset are *alike* and patterns across subsets are *not alike*, according to some criterion. The two main approaches to partitional clustering are called parametric and non-parametric.

* THIS WORK WAS PARTIALLY SUPPORTED BY NSF GRANTS ECS-9900394 AND EIA-0135946

† robertj@cnel.ufl.edu, Phone (+1) 352-392-2682, Fax. (+1) 352-392-0044

In parametric approaches some knowledge about the clusters' structure is assumed. The most famous and popular such method is McQueen's K -means algorithm [12], which implicitly assumes Gaussian cluster distributions. It minimizes a sum-of-squares cost function, equivalent to variance minimization, and thus fails if the cluster distributions are not hyper-elliptical. Often, however, there is no a-priori knowledge about the data structure. In such cases it is more natural to adopt non-parametric approaches, which make no model assumptions, such as e.g. hierarchical clustering [9]. Implicitly, usually these methods also rely on a minimum variance criterion as the clustering metric [16].

In order to capture data structure beyond second order statistics, information-theoretic clustering metrics appear as an appealing alternative. Information theory has been used in clustering by Watanabe [20], and by many other researchers, e.g. [19] and [16]. The major problem of clustering based on information-theoretic measures has been the difficulty to evaluate the metric without imposing unrealistic parametric assumptions about the data distributions.

Recently, Principe et al. [14] avoided this problem by combining a non-parametric Parzen density estimator with Renyi's [15] definition of entropy, resulting in an entropy estimator with a very interesting physical interpretation as a potential energy field. The derivative of the potential field with respect to a particle can be regarded as a force acting on that particle. This force is referred to as an Information Force, and will generally point toward a cluster. Jenssen et al. [10], utilized these forces to build directed trees, each corresponding to a cluster. In this paper we develop this method further, and use it to create an initial clustering of the data set. Renyi's entropy was also used in a somewhat similar manner in [11] for clustering.

Another information-theoretic quantity, which lends itself nicely to estimation via the Parzen window method, was proposed by Gokcay and Principe [8]. They proposed a divergence measure between pdf's, from which they defined the Cluster Evaluation Function (CEF). The CEF was utilized for clustering by minimizing it, with positive results. We re-visit the proposed pdf divergence measure, and define a new cluster evaluation function which we refer to as the Information Cut. The reason for this name comes from the deep connection of our information-theoretic measure to the graph-theoretic notion of a cut. The cut has been used rather successfully, though with some known difficulties, as a cost function in clustering and image segmentation.

A graph consists of the node set C , with a (symmetric) similarity weight G_{ij} between nodes i and j . A graph can be partitioned into two disjoint sets C_1 and C_2 simply by removing edges between the two parts. The degree of similarity between these two pieces can be computed as a total weight of the edges that have been removed. This quantity is called the cut [18]

$$\text{CUT}(C_1, C_2) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N M_{ij} G_{ij}, \quad (1)$$

where N is the total number of nodes in the graph, and M_{ij} is a membership function that equals one if and only if node i and j belong to different sub graphs. The cut has been proposed by Wu and Leahy [21] as a cost function for clustering, by minimizing it. However, as was also noted in [21], the minimum cut criteria favors a skewed cut, isolating one or a few nodes in C_1 , and the rest in C_2 . To compensate for this fact, a number of rather heuristically motivated improvements to the cut-cost have been proposed. All of these methods aim to minimize the similarity between sub graphs, i.e. the cut, while the similarity within each sub graph is maximized. Specifically, we mention the normalized cut proposed by Shi and Malik [18], Ding et al.'s min-max cut [3], the typical cut due to Gdalyahu et al. [6] and the BCut proposed by Scanlon and Deo [17].

In the next section we derive the Information Cut, and show how it can be considered a theoretically sound graph-theoretic cost function for clustering. In section 3 we define the Information Forces, following the outline given in [14]. In section 4 we propose a novel information-based clustering approach combining the Information Forces and the Information Cut. Some clustering experiments are presented in section 5. Finally, in section 6, we make our concluding remarks.

INFORMATION CUT

We define the following divergence measure between the two pdf's $p(\mathbf{x})$ and $q(\mathbf{x})$;

$$D(p, q) = -\log \left(\frac{\int p(\mathbf{x})q(\mathbf{x}) d\mathbf{x}}{\sqrt{\int p^2(\mathbf{x}) d\mathbf{x} \int q^2(\mathbf{x}) d\mathbf{x}}} \right). \quad (2)$$

This measure is always positive, it vanishes if and only if $p(\mathbf{x}) = q(\mathbf{x})$ and it is symmetric. Maximizing the divergence between $p(\mathbf{x})$ and $q(\mathbf{x})$ is equivalent to minimizing the argument of the logarithm. It is interesting to note that $D(p, q)$ can be interpreted as the Cauchy-Schwartz distance between $p(\mathbf{x})$ and $q(\mathbf{x})$ [14].

Assume that we estimate $p(\mathbf{x})$ based on the data points in $C_1 = \{\mathbf{x}_i\}$, $i = 1, \dots, N_p$, and $q(\mathbf{x})$ based on $C_2 = \{\mathbf{x}_j\}$, $j = 1, \dots, N_q$. By the Parzen [13] method

$$\hat{p}(\mathbf{x}) = \frac{1}{N_p} \sum_{i=1}^{N_p} G(\mathbf{x} - \mathbf{x}_i, \sigma^2 \mathbf{I}), \quad \hat{q}(\mathbf{x}) = \frac{1}{N_q} \sum_{j=1}^{N_q} G(\mathbf{x} - \mathbf{x}_j, \sigma^2 \mathbf{I}), \quad (3)$$

where we have used a symmetric Gaussian kernel, $G(\mathbf{x}, \mathbf{\Sigma})$, where the covariance matrix, $\mathbf{\Sigma}$, is given by $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$.

By substituting (3) into (2), and utilizing the properties of the Gaussian kernel, we define the Information Cut (IC) as the argument of the logarithm

of $D(p, q)$, given by

$$\begin{aligned}
\text{IC} &= \frac{\frac{1}{N_p N_q} \sum_{i=1}^{N_p} \sum_{j=1}^{N_q} G_{ij, 2\sigma^2 \mathbf{I}}}{\sqrt{\frac{1}{N_p^2} \sum_{i=1}^{N_p} \sum_{i'=1}^{N_p} G_{ii', 2\sigma^2 \mathbf{I}} \frac{1}{N_q^2} \sum_{j=1}^{N_q} \sum_{j'=1}^{N_q} G_{jj', 2\sigma^2 \mathbf{I}}}} \\
&= \frac{\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N M_{ij} G_{ij, 2\sigma^2 \mathbf{I}}}{\sqrt{\sum_{i=1}^N \sum_{j=1}^N M_{C_{1ij}} G_{ij, 2\sigma^2 \mathbf{I}} \sum_{i=1}^N \sum_{j=1}^N M_{C_{2ij}} G_{ij, 2\sigma^2 \mathbf{I}}}}, \quad (4)
\end{aligned}$$

where $N = N_p + N_q$, $G_{ij, 2\sigma^2 \mathbf{I}} = G(\mathbf{x}_i - \mathbf{x}_j, 2\sigma^2 \mathbf{I})$ and $M_{C_{1ij}}$ and $M_{C_{2ij}}$ are membership functions that equal one if and only if $\mathbf{x}_i$ and $\mathbf{x}_j$ both belong to C_1 , or C_2 , respectively.

If $\mathbf{x}_i$ and $\mathbf{x}_j$ are considered nodes in the set C , it is clear that $G_{ij, 2\sigma^2 \mathbf{I}}$ in fact is the similarity weight between the two nodes. Hence, when partitioning the set C into the sub sets, or clusters, C_1 and C_2 , the numerator of the Information Cut is exactly the traditional cut known from graph theory, defined in (1). Very interestingly, our information-theoretic approach reveals that the cut should be normalized by the square root of the product of the sum of the total weights in C_1 and C_2 , respectively, which is the quantity in the denominator of (4).

There is also another interesting interpretation of the Information Cut. Actually, the Gaussian kernel used in the Parzen pdf estimation can be regarded a Mercer kernel [7], [1], performing a nonlinear data transformation into some high dimensional feature space, which increases the probability of linear separability of the clusters in the transformed space. Inner products in the feature space are implicitly computed in the input space by use of the kernel. This means that $G_{ij, 2\sigma^2 \mathbf{I}} = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, where $\Phi(\cdot)$ denotes the mapping from input space to feature space. Minimizing the Information Cut is equivalent to minimizing the sum of inner products across clusters in feature space, while at the same time maximizing the product of the sums of inner products within clusters. We mention that Cristianini et al. [2] used the traditional cut-cost in spectral kernel methods for clustering.

In the case of multiple clusters, C_k , $k = 1, \dots, K$, we extend the previous definition of the Information Cut as follows;

$$\text{IC} = \frac{\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N M_{ij} G_{ij, 2\sigma^2 \mathbf{I}}}{\sqrt{\prod_{k=1}^K \sum_{i=1}^N \sum_{j=1}^N M_{C_{kij}} G_{ij, 2\sigma^2 \mathbf{I}}}}. \quad (5)$$

The CEF [8] was derived in a similar manner as the Information Cut. However, in that case, only the numerator of the argument of the logarithm in the expression for $D(p, q)$ was considered. In fact, the CEF and the BCut [17] are equal up to a factor of two.

INFORMATION FORCES

Another information-theoretic quantity which lends itself nicely to non-parametric estimation is Renyi's entropy. For $\alpha > 0$, $\alpha \neq 1$, Renyi's entropy for a stochastic variable $\mathbf{x}$ with pdf $f(\mathbf{x})$ is given by [15]

$$H_R(\mathbf{x}) = \frac{1}{1-\alpha} \log \int f^\alpha(\mathbf{x}) d\mathbf{x} = -\log \int f^2(\mathbf{x}) d\mathbf{x}, \quad (6)$$

for the specific choice of $\alpha = 2$ [14]. This is called Renyi's quadratic entropy.

Renyi's quadratic entropy can also easily be estimated directly from data by the use of Parzen pdf estimation. When we plug the pdf estimate into (6) we obtain;

$$H_R(\mathbf{x}) = -\log \left(\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G_{ij, 2\sigma^2 \mathbf{I}} \right) = -\log V_R(\mathbf{x}), \quad (7)$$

where N is the total number of data points the pdf estimate is based on.

Regarding the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ as physical particles, we can regard $V_{ij} = G_{ij, 2\sigma^2 \mathbf{I}}$ as an interaction law between particles imposed by the Gaussian kernel. The sum of interactions on the i 'th particle is $V_i = \sum_j V_{ij}$. The sum of all pairs of interactions, given by $V_R(\mathbf{x})$ in (7), can now be regarded as an overall potential energy of the data set. This potential energy is called the Information Potential (IP). It is interesting at this point to note that the IP is nothing but the normalized sum of all weights in a graph. Because of symmetry, weights between different nodes are counted twice. It is also interesting to note that the numerator of the Information Cut can be considered the Information Potential between sub groupings, while in the normalizing denominator the product of the Information Potentials within each sub grouping appears.

Just as in mechanics, the force acting on particle $\mathbf{x}_i$ is given by the derivative of the potential field with respect to the particle;

$$\mathbf{F}_i = \frac{\partial}{\partial \mathbf{x}_i} V_R(\mathbf{x}) = -\frac{1}{N^2 \sigma^2} \sum_{j=1}^N V_{ij} \mathbf{d}_{ij}, \quad (8)$$

where $\mathbf{d}_{ij} = \mathbf{x}_i - \mathbf{x}_j$. This is the net effect of the IP on particle $\mathbf{x}_i$, and will be called an Information Force. Jenssen et al. [10] showed that if σ is chosen such that the pdf estimate is relatively accurate, each Information Force points toward a cluster. The Information Forces can also be interpreted in light of inner products in a high dimensional feature space.

INFORMATION-BASED CLUSTERING APPROACH

We develop a novel clustering approach, utilizing the Information Forces to obtain an initial clustering of the data set. The initial clusters correspond

to directed trees [5], which we create according to the direction of the Information Forces and the similarity weight between nodes. Subsequently we use the Information Cut as a cluster evaluation function, merging clusters pairwise until the predefined number of clusters is reached. Note that e.g. Girolami [7] has shown that kernel based clustering can also provide us with a tool for determining the number of clusters in the data set.

In a directed tree each node i initiates (points to) another node j , which is called the predecessor of i . Only one node does not have a predecessor, and this node is called the root.

Our procedure for creating directed trees is very simple. We examine every data point $\mathbf{x}_i$, $i = 1, \dots, N$, one at a time, where $\mathbf{x}_i$ corresponds to node i in the final tree. For node i we determine whether it has a predecessor j , or whether it is a root, based on the following: Node j is the predecessor of node i if it satisfies

$$\max_{j=1, \dots, N} \cos \angle(\mathbf{F}_i, \mathbf{x}_j - \mathbf{x}_i) G_{ij, \sigma^2 \mathbf{I}}, \quad j \neq i \quad (9)$$

under the following constraint;

- Node j can not be one of i 's children.

When examining node i we sort all $\cos \angle(\mathbf{F}_i, \mathbf{x}_j - \mathbf{x}_i) G_{ij, \sigma^2 \mathbf{I}}$, $j = 1, \dots, N, j \neq i$, in descending order, and store the result in the vector **pre**, corresponding to possible predecessors of i . If node j corresponds to the first element of **pre**, but actually is one of i 's children, we know that node i is possibly a root. However, we consider the node k corresponding to the second element (and third, and so on) of **pre** as a predecessor of node i , but only if $\mathbf{F}_k$ points toward $\mathbf{F}_i$ within a 45° sector (rather heuristically chosen), and $\cos \angle(\mathbf{F}_i, \mathbf{x}_k - \mathbf{x}_i) > 0$. If these requirements are not met, it is likely that node k belongs to a different cluster than node i . When the first node not being one of i 's children, and not satisfying the requirements above is found, the search for a predecessor is terminated, and node i is declared a root.

Clustering with the Information Forces alone has a tendency to create a few spurious clusters, even if the value of σ corresponds to a relatively accurate Parzen pdf estimate. We therefore merge all initial clusters with less than 10 members to the nearest cluster, with respect to cluster means. Subsequently, we evaluate the Information Cut for all combinations of pairwise cluster mergings, and eventually merge those two clusters that result in the minimum Information Cut value. This procedure is repeated until the predefined number of clusters is reached.

RESULTS

In all experiments the data have been normalized feature-by-feature to have a range $[-1, 1]$. This is mainly done in order for the results to be comparable to previous results e.g. [10],[11] and [8].

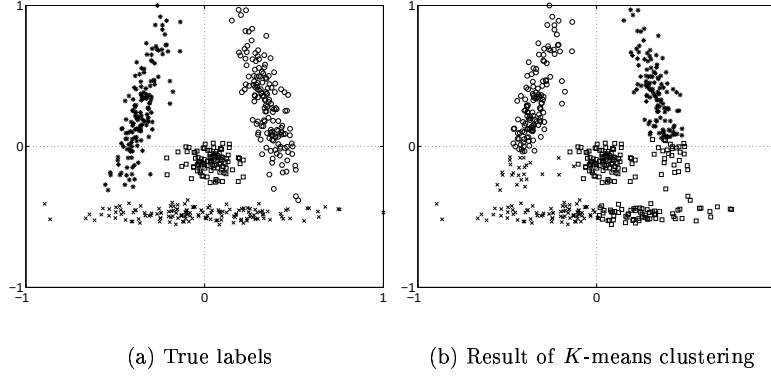


Figure 1: Artificial data set.

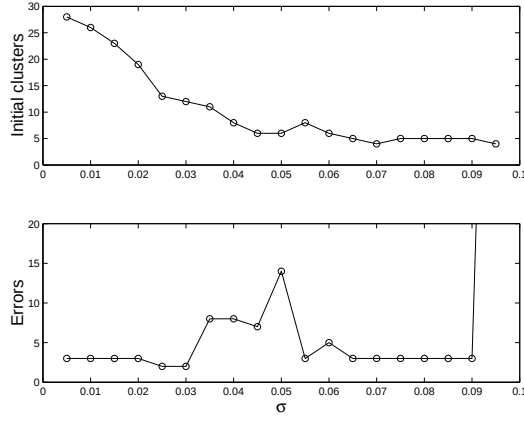


Figure 2: Result of proposed information-theoretic clustering method.

We first consider the data set shown in Fig. 1 (a) where data patterns belonging to the same cluster have been labeled with the same symbol. To show that K -means, based on second order statistics, is incapable of producing a correct labeling of this data set, we ran it 10 times, and display in Fig. 1 (b) the best result.

In the upper panel of Fig. 2 we show the number of initial clusters produced by the Information Force directed tree method, after merging of spurious clusters, for a wide range of σ -values. The lower panel show the resulting errors after Information Cut merging. It is clear that the result is very satisfying, possibly except the middle range of σ -values. Jenssen et al. [10] studied the same data set, and noted that the Parzen pdf estimate corresponding to σ in the middle range and lower, gets increasingly crude and noisy. This is also reflected in the Information Forces, and some problems occur in the area to the lower right of the data set, where there are some overlap. But for even

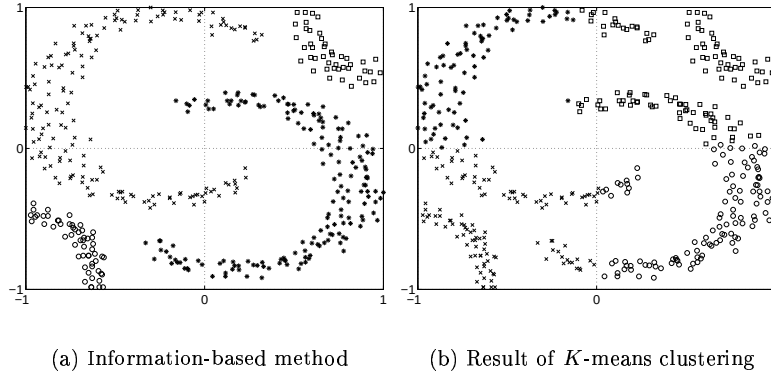


Figure 3: Highly irregular clusters.

smaller σ values, fewer and fewer patterns interact with each other, resulting in many relatively small clusters being formed, but where almost none of the resulting initial clusters spread over the true clusters. In this situation the Information Cut evaluates the clusters perfectly, and is able to recover the true clusters very well. It is worth mentioning that σ 's in this range do not at all correspond to smooth pdf estimates. For σ -values in the upper range ($\approx 0.07 - 0.08$) the Information Force clustering does the bulk of the job, but still the Information Cut merges correctly the few clusters that need to be merged. However, the smoothing effect resulting from a large kernel size increasingly starts to dominate, and for $\sigma \geq 0.095$, the Information Forces no longer sees four clusters, but rather three, and the method brakes down.

We also test our method on a data set consisting of highly irregular clusters. For very small σ -values, up to $\sigma \approx 0.1$, we obtain a perfect clustering. For example, in Fig. 3 (a) we show the result obtained for $\sigma = 0.05$. For $\sigma > 0.1$ the resulting clusters tend to spread over the true cluster boundaries. Not surprisingly, K -means fails completely, as shown in Fig. 3 (b).

For data sets consisting of clusters with a large degree of overlap we expect our method to encounter problems, mainly because the Information Forces may no longer point toward distinct cluster centers. Therefore, when clustering the well known four-dimensional IRIS¹ data set, the range of σ 's for which we obtain reasonable results is more narrow. However, for $0.028 \leq \sigma \leq 0.046$, we obtain satisfactory results yielding six errors. The errors are the same over the whole range of σ -values, and the problem stems from the overlap of two of the three true clusters. Our result is not as good as e.g. [7], [16], (three errors), but compares favorably with those reported in e.g. [8], [1], [19], [4] and several others.

We also cluster the 13-dimensional WINE data set. The results we obtain can not compare to those reported in [16], [7] (four errors). Over a wide range of σ -values, we obtain in the order of 10 errors. This compares to [4].

¹IRIS and WINE data sets extracted from the UCI repository, University of California, Irvine.

CONCLUSION

We have introduced the Information Cut, and combined it with the Information Forces in a novel non-parametric information-theoretic clustering algorithm. The main advantage of our clustering approach is that the underlying clustering metric is based on entropy, both between sub groups, and within sub groups. Entropy is a quantity that conveys information about the shape of probability distributions, and not only variance, which many traditional clustering algorithms, e.g. K -means rely on. This enables us to cluster data sets consisting of elongated and highly irregular clusters.

However, for data sets consisting of clusters with a high degree of overlap, our method encounter increasing difficulty because the Information Forces no longer point to distinct cluster centers.

To calculate the Information Forces and to find the predecessor of each node, we need to run through the data set once. Each time the distance from a data pattern to all other patterns must be calculated and evaluated by the Gaussian kernel. This is in itself an $O(N)$ operation, resulting in an overall $O(N^2)$ operation. If memory requirements allow for it, all the similarity weights can be stored in a matrix so that they need not be recalculated when we utilize the Information Cut for merging. Thus, in the merging phase, all calculations are simple matrix manipulations. Naturally, the Information Cut merging operation will take longer time the more clusters need to be merged.

At present, the major problem with our clustering algorithm is how to choose the kernel size σ . This is a problem encountered in all kernel-based methods, both supervised and unsupervised. However, we have shown that it is most critical to avoid choosing σ too large, since the Information Cut rather successfully merges small clusters resulting from small σ . Important topics for future research include developing an automatic procedure to determine σ such that the corresponding Parzen pdf estimate is relatively accurate, or at least not too smooth.

REFERENCES

- [1] A. Ben-Hur, D. Horn, H. T. Siegelmann and V. Vapnik, "Support Vector Clustering," **Journal of Machine Learning Research**, vol. 2, pp. 125–137, 2001.
- [2] N. Cristianini, J. Shawe-Taylor and J. Kandola, "Spectral Kernel Methods for Clustering," in **Advances in Neural Information Processing Systems**, 2002, vol. 1, pp. 649–655.
- [3] C. H. Q. Ding, X. He, H. Zha, M. Gu and H. D. Simon, "A Min-max Cut Algorithm for Graph Partitioning and Data Clustering," in **IEEE Int. Conf. on Data Mining**, 2001, pp. 107–114.
- [4] T. Eltoft and R. J. P. deFigueiredo, "A New Neural Network for Cluster-Detection-and-Labeling," **IEEE Transactions on Neural Networks**, vol. 9, no. 5, pp. 1021–1035, 1998.

- [5] K. Fukunaga, **Introduction to Statistical Pattern Recognition**, New York: Academic Press, 1990.
- [6] Y. Gdalyahu, D. Weinshall and M. Werman, "Self-Organization in Vision: Stochastic Clustering for Image Segmentation, Perceptual Grouping, and Image Database Organization," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 23, no. 10, pp. 1053–1074, 2001.
- [7] M. Girolami, "Mercer Kernel-Based Clustering in Feature Space," **IEEE Transactions on Neural Networks**, vol. 13, no. 3, pp. 780–784, 2002.
- [8] E. Gokcay and J. Principe, "Information Theoretic Clustering," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 24, no. 2, pp. 158–170, 2002.
- [9] A. K. Jain and R. C. Dubes, **Algorithms for Clustering Data**, Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [10] R. Jenssen, D. Erdogmus, K. E. Hild, J. C. Principe and T. Eltoft, "Information Force Clustering using Directed Trees," in **Accepted to International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition**, Lisbon, Portugal, 2003.
- [11] R. Jenssen, K. E. Hild, D. Erdogmus, J. C. Principe and T. Eltoft, "Clustering using Renyi's Entropy," in **Accepted to International Joint Conference on Neural Networks**, Portland, Oregon, USA, 2003.
- [12] J. McQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in **Fifth Berkeley Symposium on Mathematical Statistics and Probability**, 1967, pp. 281–297.
- [13] E. Parzen, "On the Estimation of a Probability Density Function and the Mode," **Ann. Math. Stat.**, vol. 32, pp. 1065–1076, 1962.
- [14] J. Principe, D. Xu and J. Fisher, "Information Theoretic Learning," in **Unsupervised Adaptive Filtering**, S. Haykin (Ed.), John Wiley & Sons, 2000, vol. I, Chapter 7.
- [15] A. Renyi, "On Measures of Entropy and Information," in **Fourth Berkeley Symposium on Mathematical Statistics and Probability**, 1960, pp. 547–561.
- [16] S. J. Roberts, R. Everson and I. Rezek, "Minimum Entropy Data Partitioning," in **Ninth International Conference on Artificial Neural Networks**, 1999, vol. 2, pp. 844–849.
- [17] J. Scanlon and N. Deo, "Graph-Theoretic Algorithms for Image Segmentation," in **IEEE International Symposium on Circuits and Systems**, Orlando, Florida, 1999, pp. VI141–144.
- [18] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 22, no. 8, pp. 888–905, 2000.
- [19] N. Tishby and N. Slonim, "Data Clustering by Markovian Relaxation and the Information Bottleneck Method," in **Advances in Neural Information Processing Systems**, Denver, USA, 2000, vol. 13, pp. 640–646.
- [20] S. Watanabe, **Pattern Recognition: Human and Mechanical**, John Wiley & sons, 1985.
- [21] Z. Wu and R. Leahy, "An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Applications to Image Segmentation," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 15, no. 11, pp. 1101–1113, 1993.