CAUCHY-SCHWARTZ PDF DIVERGENCE MEASURE FOR NON-PARAMETRIC CLUSTERING

Robert Jenssen[1,2], *Jose C. Principe*[2] *and Torbjørn Eltoft*[1]

[1]Department of Physics
University of Tromsø, Norway

[2]Computational NeuroEngineering Laboratory
University of Florida, USA

ABSTRACT

We propose a new cost function for clustering based on the Cauchy-Schwartz (CS) inequality. This cost function is obtained by replacing inner products between vectors with inner products between probability density functions (pdf's), in the CS inequality expression. Combined with Parzen pdf estimation, this provides us with a non-parametric information-theoretic cluster evaluation function that we refer to as the CS divergence. We propose a novel method for maximization of the CS divergence for clustering, and present results on both artificial data and real data.

1. INTRODUCTION

In exploratory data analysis it is often desirable to perform an unsupervised classification 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. This problem is known as clustering [1]. Clustering has become an important tool in areas such as data mining [2], image segmentation [3], signal compression [4] and machine learning [5].

The two main approaches to clustering can be divided into the parametric and the non-parametric methods. In parametric methods some knowledge about the clusters' structure is assumed. The most famous and popular such method is McQueen's K-means algorithm [6], 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 hyperelliptical. 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. single-link or complete-link hierarchical clustering [1]. Implicitly, usually these methods also rely on a minimum variance criterion as the clustering metric [7].

In order to capture data structure beyond second order statistics, information-theoretic clustering metrics, such as entropy, mutual information and Kullback-Leibler divergence [8], appear as an appealing alternative. Information theory has been used in clustering by Watanabe [9], and by several other researchers, e.g. [10, 11, 7]. 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. [12] proposed a pdf divergence measure that lends itself nicely to non-parametric estimation via Parzen windowing [13]. This pdf divergence measure is based on the Cauchy-Schwartz inequality between two vectors. Combined with a Gaussian kernel in Parzen pdf estimation, the Cauchy-Schwartz pdf divergence was utilized for blind source separation and pose estimation in synthetic aperture radar imagery [12]. Later, Gokcay and Principe [14] used a somewhat similar measure for clustering, with positive results.

In this paper we show that the Cauchy-Schwartz pdf divergence measure can be utilized as a cost function for non-parametric clustering. The remainder of the paper is organized as follows. In the next section we define the Cauchy-Schwartz pdf divergence measure, and show how it can be estimated directly from the available data in a non-parametric fashion. In section 3 we propose a novel method for maximization of this cost function for clustering. In section 4 we illustrate the performance of the proposed method using some artificial data and some real data. Finally, in section 5 we make our concluding remarks.

This work was partially supported by grants ECS-0300340 and EIA-0135946

R. Jenssen was supported by a scholarship from the University of Tromsø

Corresponding author is R. Jenssen. Phone: (+1) 352-392-2682, Fax: (+1) 352-392-0044, Email: robertj@phys.uit.no

2. CAUCHY-SCHWARTZ PDF DIVERGENCE MEASURE

Based on the Cauchy-Schwartz inequality; $||\mathbf{x}||^2\ ||\mathbf{y}||^2 \geq (\mathbf{x}^T\mathbf{y})^2$, the following holds;

$$-\log \frac{\mathbf{x}^T\mathbf{y}}{\sqrt{||\mathbf{x}||^2\ ||\mathbf{y}||^2}} \geq 0. \tag{1}$$

In the spirit of Eq. (1), we define the following divergence measure between the two pdf's $p(\mathbf{x})$ and $q(\mathbf{x})$ [12];

$$D_{CS}(p,q) = -\log \frac{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}{\sqrt{\int p^2(\mathbf{x})d\mathbf{x}\ \int q^2(\mathbf{x})d\mathbf{x}}}. \tag{2}$$

The logarithm is included to make the notation consistent with that of other divergence measures. This measure can be regarded as an approximation of the Kullback-Leibler divergence between the two pdf's. It 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.

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

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

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

Now, we define the membership function M_{ij}, which equals one iff the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different clusters, and zero if not. Furthermore, we define the membership functions $M_{C_{1_{ij}}}$ and $M_{C_{2_{ij}}}$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ both belong to C_1, then $M_{C_{1_{ij}}} = 1$, but zero if not. Likewise for $M_{C_{2_{ij}}}$ with respect to C_2. By substituting (3) into (2), and utilizing the properties of the Gaussian kernel [12], we can estimate $D_{CS}(p,q) = -\log V_{CS}(p,q)$ in a non-parametric fashion, where;

$$\begin{aligned}&\hat{V}_{CS}(p,q) = \\ &= \frac{\frac{1}{N_pN_q}\sum_{i,j}^{N_p,N_q} G_{ij,2\sigma^2\mathbf{I}}}{\sqrt{\frac{1}{N_p^2}\sum_{i,i'}^{N_p,N_p} G_{ii',2\sigma^2\mathbf{I}}\ \frac{1}{N_q^2}\sum_{j,j'}^{N_q,N_q} G_{jj',2\sigma^2\mathbf{I}}}}\\ &= \frac{\frac{1}{2}\sum_{i,j}^{N,N} M_{ij}G_{ij,2\sigma^2\mathbf{I}}}{\sqrt{\sum_{i,j}^{N,N} M_{C_{1_{ij}}}G_{ij,2\sigma^2\mathbf{I}}\ \sum_{i,j}^{N,N} M_{C_{2_{ij}}}G_{ij,2\sigma^2\mathbf{I}}}},\end{aligned} \tag{4}$$

where $\sum_{i,j}^{N,N} G_{ij,2\sigma^2\mathbf{I}} = \sum_{i=1}^{N}\sum_{j=1}^{N} G_{ij,2\sigma^2\mathbf{I}}$, $N = N_p + N_q$ and $G_{ij,2\sigma^2\mathbf{I}} = G(\mathbf{x}_i - \mathbf{x}_j, 2\sigma^2\mathbf{I})$.

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

$$\hat{V}_{CS}(p,q) = \frac{\frac{1}{2}\sum_{i,j}^{N,N} M_{ij}G_{ij,2\sigma^2\mathbf{I}}}{\sqrt{\prod_{k=1}^{K}\sum_{i,j}^{N,N} M_{C_{k_{ij}}}G_{ij,2\sigma^2\mathbf{I}}}}. \tag{5}$$

At this point we note a particularly interesting feature of the Cauchy-Schwartz divergence. Based on Eq. (5), it is easily shown that $\hat{D}_{CS}(p,q)$ is in fact an estimate of Renyi's quadratic entropy [12] calculated between samples of different clusters, subtracted (half) the sum of Renyi's quadratic entropy of each individual cluster. This means that in order for $D_{CS}(p,q)$ to be large, the sum of the entropies of the individual clusters must be small, while at the same time the entropy across the clusters must be large. This makes perfect sense.

There is also another interesting interpretation of the Cauchy-Schwartz divergence. Actually, the Gaussian kernel used in the Parzen pdf estimation can be regarded a Mercer kernel [15], 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 $V_{CS}(p,q)$ 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.

In conclusion, Eq. (5) provides us with an information-theoretic cluster evaluation function, capable of capturing data structure beyond mere second order statistics, as many traditional clustering algorithms are restricted to.

3. MAXIMIZING THE DIVERGENCE

In this section we propose a novel method for maximization of the Cauchy-Schwartz pdf divergence. The method we propose here has close resemblance to the approach taken by Jenssen et al. [10] in their entropy-based algorithm. The main idea is to “seed” a number of small initial clusters in the data set, grow the clusters until all patterns have been labeled, and then re-cluster the members of the “worst” cluster, thus reducing the number of clusters by one. This procedure is repeated until the predefined number of clusters is reached.

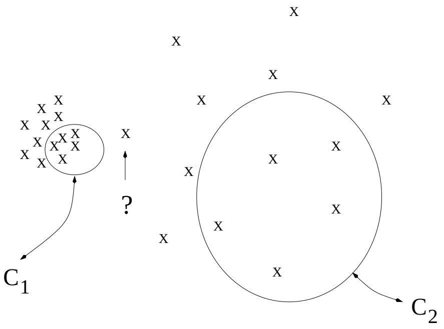


Figure 1: Assigning a data pattern to a cluster.

3.1. Growing Clusters

Consider the situation depicted in Fig. 1. A set of patterns, or feature vectors, is distributed in feature space. Initially a subset of the feature vectors have been assigned to cluster C_1 or C_2. These are shown as the encircled points. The problem of clustering is now to decide whether a new pattern $\mathbf{x}$ (pointed to by the question mark) should be assigned to C_1 or C_2. Our solution to this problem is very simple: Assign $\mathbf{x}$ to C_1 if $D_{CS}(C_1+\mathbf{x}, C_2) > D_{CS}(C_1, C_2+\mathbf{x})$, and to C_2 if the opposite is true.

Hence, in the general case of having initial clusters $C_k,\ k = 1,\ldots,K$, assign $\mathbf{x}$ to cluster C_i if

$$\max_i D_{CS}(C_1,\ldots,C_i+\mathbf{x},\ldots,C_K), \qquad (6)$$

for $i = 1,\ldots,K$.

This approach to clustering is both intuitive and simple. However, at this point two questions arise: How to initially cluster a subset of the data? And, how to decide which pattern to be clustered next?

Initially we "seed" K_{init} clusters in the data set. This is done by randomly selecting K_{init} "seed" patterns, one at a time. The first "seed" pattern and it's $N_{\text{init}}-1$ nearest neighbors constitute the first cluster. Thereafter a new pattern among the unlabeled patterns is randomly selected, which together with it's $N_{\text{init}}-1$ closest unlabeled neighbors constitute the second cluster. This process is repeated until K_{init} clusters have been formed.

The next pattern to be clustered can be selected in several ways.

- Randomly among the unlabeled patterns - this has the disadvantage that early clustering of points far from the initial clusters can make the clustering process unstable.

- As the unlabeled pattern closest to a cluster prototype - this approach makes the clustering more stable. If we choose to have just one cluster prototype per cluster, it is typically the cluster mean. At the other extreme, all patterns in a cluster are regarded prototypes.

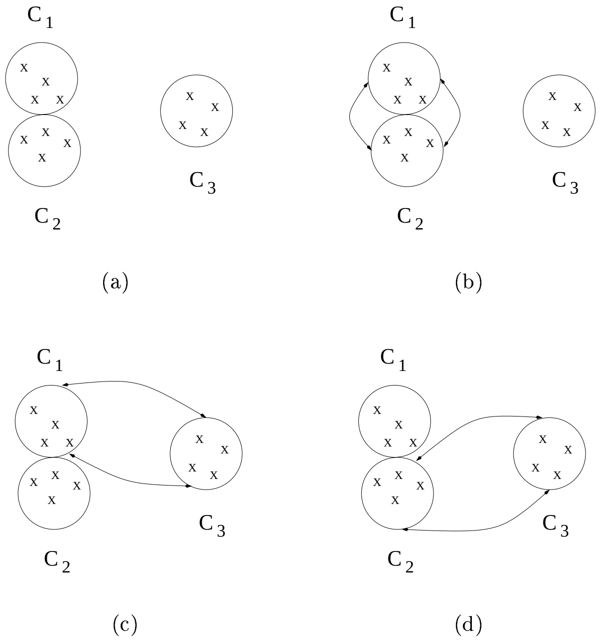


Figure 2: Identifying the "worst cluster" using the Cauchy-Schwartz pdf divergence. In this example C_1 or C_2 is the "worst cluster."

3.2. Cluster Evaluation

In Fig. 2 we have illustrated how the CS divergence can be used to identify the "worst cluster", whose members subsequently are re-clustered. In (a) a data set has been partitioned into three clusters. We proceed by eliminating one cluster at a time, and calculate the CS divergence based on the remaining clusters in each case. To be more specific, by eliminating, we mean that the members of the eliminated cluster are considered unlabeled, not contributing to the value of the CS divergence. For example, in (b) C_3 is eliminated, and the CS divergence based on C_1 and C_2 alone is calculated, as indicated by the arrows. Likewise in (c) and (d).

The "worst cluster" is now selected as the cluster that when eliminated, results in the largest CS divergence based on the remaining clusters, because this means that the remaining clusters are the most separated clusters. In the situation depicted in Fig. 2, this method results in either C_1 or C_2 being identified as the "worst cluster".

3.3. Complexity and instability

Usually the CS divergence clustering algorithm performs better by choosing the next pattern to be clustered as the unlabeled pattern closest to a cluster prototype. This clearly increases the complexity of the algorithm compared to random selection. In our implementation of the algorithm, the total complexity is always less than $O(N^2)$ at any iteration.

Because the initial clusters are "seeded" randomly, the clustering result can differ when clustering the same data set several times. The CS divergence calculated for the final clustering in each case can provide us with a good indication of which random initialization provided the best result.

4. PERFORMANCE STUDIES

In this section we test the performance of our novel algorithm on four data sets, two artificially created and two real. In all experiments the data have been normalized to have a range $[-1, 1]$. This is mainly done in order to have some control over the parameter σ. It doesn't affect the structure of the data.

The first data set is shown in Fig. 3. This data set consists of one spherical cluster and three elongated clusters, with a total of 550 patterns. 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. 3 (b) the best result. The clusters are represented by different symbols. Fig. 3 (a) shows the best result (but also typical result) obtained by the CS method for $\sigma = 0.06$. In this case we "seed" $K_{init} = 20$ clusters, each with $N_{init} = 10$ members. The next pattern to be clustered is the one closest to some labeled pattern. The result is very satisfying, yielding only two errors. The thick lines identifies the erroneously labeled patterns, and indicates that they actually belong to the cluster marked by squares. Our experiments show that these two patterns are always wrongly labeled by the CS method. However, this is to be expected, since by visual examination of the data set, a human would probably also assign these two patterns to the lower horizontal cluster. Next, we investigate the sensitivity of the method wrt. σ and K_{init}. First, we perfom an experiment where we keep K_{init} and N_{init} fixed at 20 and 10, respectively. We apply the CS method 10 times for a wide range of σ's, and show in Fig. 4 (a) the mean errors obtained. It can be seen that the CS method yields reasonable results for σ as low as 0.03, and as high as 0.17, with mean errors less than five obtained in the range $0.05 \leq \sigma \leq 0.14$. Fig. 4 (b) show the result of a similar experiment for $\sigma = 0.09$, $N_{init} = 10$, for a wide range of K_{init}. For $K_{init} < 10$, typically one or more of the ten runs result in complete failure (more than 100 errors) because of the random initialization of the initial clusters. We only show the resulting mean error when there are no complete failures. The error bars indicate the standard deviation. It can be seen that for $K_{init} \geq 17$ the CS method is very stable, and results in a mean error close to three. For $\sigma = 0.09$ we never achieved a better result than three errors. If we choose our clustering result to be the one for which the CS divergence is the smallest among the ten runs for each K_{init}, the result would be three errors for every single K_{init}, even for $K_{init} = 4$. Finally, in Fig. 5 we show the corresponding Parzen pdf estimates for three different σ's. For $\sigma = 0.03$ the pdf estimate is very crude and noisy. For $\sigma = 0.17$ the smoothing effect increasingly dominates. As shown earlier, even for these clearly inaccurate pdf estimates, the CS method performs relatively well. The pdf estimate for σ in the middle range (0.1) is also shown, for which the clustering results naturally are the best.

We also test our method on a data set consisting of highly irregular clusters. For very small σ-values (0.04), up to $\sigma = 0.12$, we obtain a perfect clustering. For example, in Fig. 6 (a) we show the result obtained for $\sigma = 0.1$, $K_{init} = 20$ and $N_{init} = 10$. For $\sigma > 0.12$ the resulting clusters tend to spread over the true cluster boundaries. Not surprisingly, K-means fails completely, as shown in Fig. 6 (b).

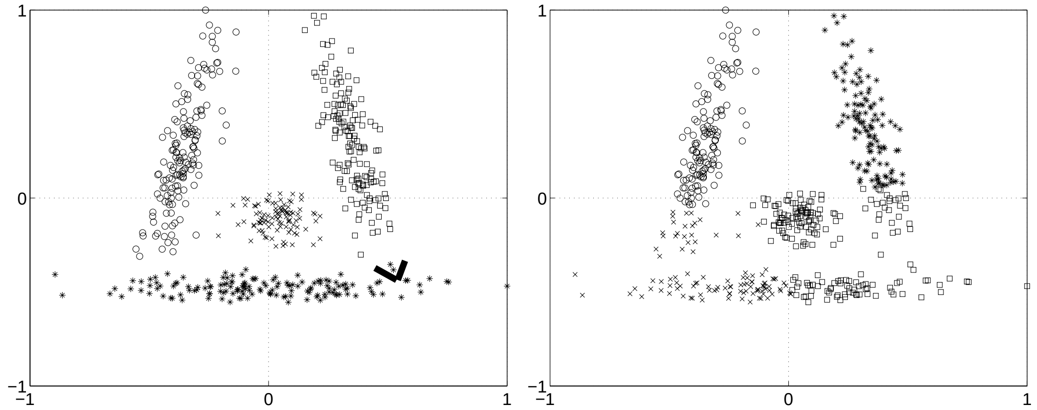
Next, we cluster the well known IRIS[1] data set. This data set consists of three classes of 50 patterns each, where each class refers to a type of iris plant. It is characterized by four numeric attributes. We run the CS algorithm ten times, and select the clustering result yielding the largest CS divergence as our final result. In this case $K_{init} = 10$ and $N_{init} = 10$. For $0.06 \leq \sigma \leq 0.3$ we obtain less than ten errors. The best result is achieved for $\sigma = 0.1$, for which we obtain only five errors. The best clustering result to our knowledge of the IRIS data set is three errors obtained by Roberts et al. [7].

Finally, we test our method on the WINE data set. This data set consists of 178 instances in a 13-dimensional feature space, where the features are found by chemical analysis of three different types of wines. We include this data set in our analysis because it shows that our algorithm is capable of performing well in a high dimensional feature space. For a wide range of σ's we obtain in the order of ten errors. The best result was obtained for $\sigma = 0.5$, yielding six errors.

5. CONCLUSION

We have introduced the Cauchy-Schwartz pdf divergence measure, and showed that it provides us with

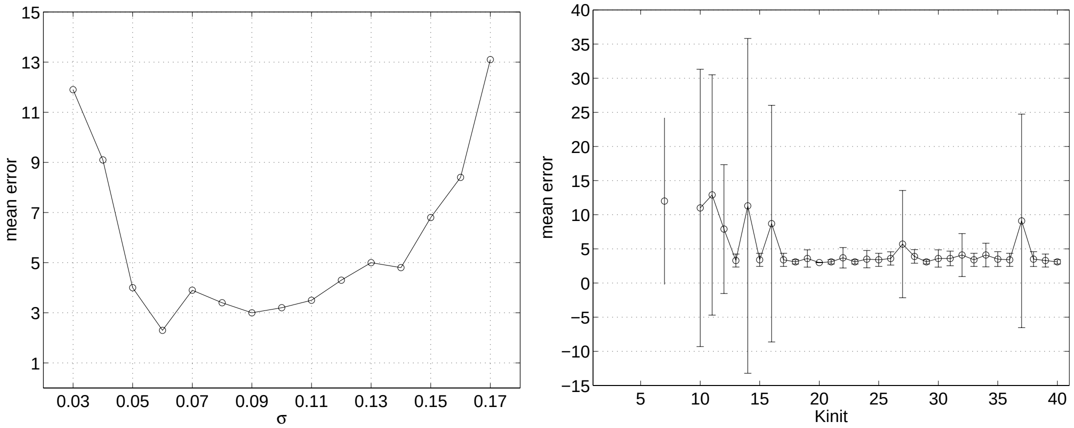
[1] IRIS and WINE data sets extracted from the UCI repository, University of California, Irvine.



(a) CS-method $\sigma = 0.06$

(b) K-means

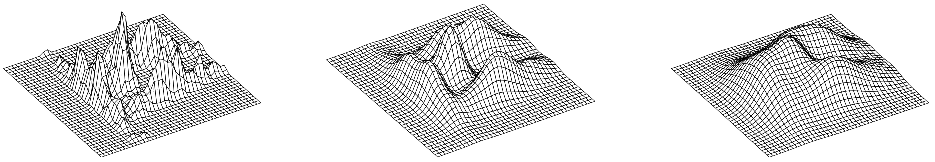
Figure 3: (a) Best result for CS-method, and (b) best result using K-means.



(a) Mean error for different σ.

(b) Mean error for different K_{init}

Figure 4: Sensitivity analysis wrt. σ and K_{init}.

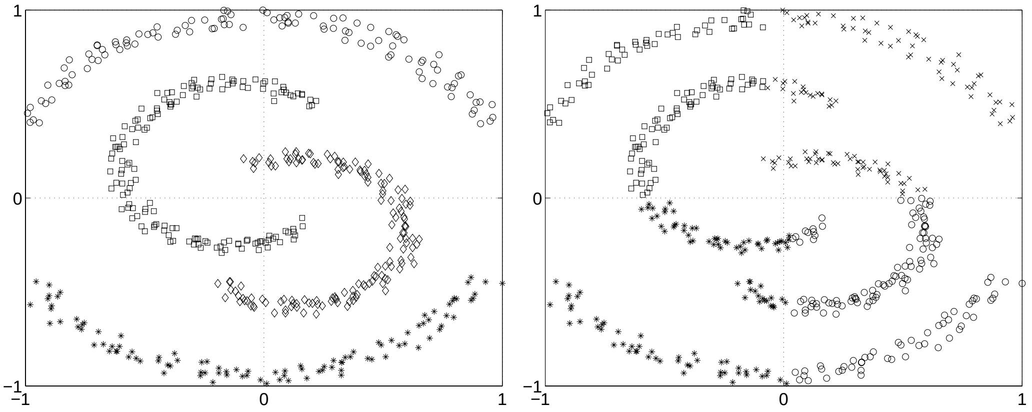


(a) $\sigma = 0.03$

(b) $\sigma = 0.1$

(c) $\sigma = 0.17$

Figure 5: Parzen estimate for different σ.



(a) CS-method $\sigma = 0.1$

(b) K-means

Figure 6: (a) Best result for CS-method, and (b) best result using K-means.

a non-parametric information-theoretic cluster evaluation function. 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.

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 the CS method is not too sensitive to the actual kernel size. Important topics for future research include developing an automatic procedure to determine σ such that the corresponding Parzen pdf estimate is relatively accurate.

6. REFERENCES

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[2] D. Judd, P. McKinley, and A. K. Jain, "Large-Scale Parallel Data Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 871–876, 1998.

[3] H. Frigui and R. Krishuapuram, "A Robust Competitive Clustering Algorithm with Applications in Computer Vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 450–465, 1999.

[4] H. M. Abbas and M. M. Fahmy, "Neural Networks for Maximum Likelihood Clustering," *Signal Processing*, vol. 36, no. 1, pp. 111–126, 1994.

[5] C. Carpineto and G. Romano, "A Lattice Conceptual Clustering System and its Application to Browsing Retrieval," *Machine Learning*, vol. 24, no. 2, pp. 96–122, 1996.

[6] J. McQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[7] 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.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & sons, 1991.

[9] S. Watanabe, *Pattern Recognition: Human and Mechanical*, John Wiley & sons, 1985.

[10] R. Jenssen, K. E. Hild, D. Erdogmus, J. C. Principe, and T. Eltoft, "Clustering using Renyi's Entropy," in *International Joint Conference on Neural Networks*, Portland, Oregon, USA, 2003, pp. 523–528.

[11] N. Tishby and N. Slonim, "Data Clustering by Markovian Relaxation and the Information Bottleneck Method," in *Advances in Neural Information Processing Systems, 13*, Denver, USA, 2000, pp. 640–646.

[12] 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.

[13] E. Parzen, "On the Estimation of a Probability Density Function and the Mode," *Ann. Math. Stat.*, vol. 32, pp. 1065–1076, 1962.

[14] E. Gokcay and J. Principe, "Information Theoretic Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 158–170, 2002.

[15] M. Girolami, "Mercer Kernel-Based Clustering in Feature Space," *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 780–784, 2002.