Proceedings of the 11th IASTED International Conference
ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING
August 29–31, 2007, Palma de Mallorca, Spain
ISBN Hardcopy: 978-0-88986-693-5 / CD: 978-0-88986-694-2

A FUZZY GAP STATISTIC FOR FUZZY C-MEANS

Christopher Sentelle
School of EECS
University of Central Florida
Orlando, FL 32816-2786, USA
email: csentelle@cfl.rr.com

Siu Lun Hong
School of EECS
University of Central Florida
Orlando, FL 32816-2786, USA
email: si061272@earthlink.net

Michael Georgiopoulos
School of EECS
University of Central Florida
Orlando, FL 32816-2786, USA
email: michaelg@mail.ucf.edu

Georgios C. Anagnostopoulos
Dept. of ECE
Florida Institute of Technology
Melbourne, FL 32901-6975, USA
email: georgio@fit.edu

ABSTRACT

The gap statistic is a statistical method for determining the number of optimal clusters for an unsupervised clustering algorithm and has been shown to outperform other cluster validity indices for the *K*-means clustering algorithm. In this paper, we assess the performance of the gap statistic when applied to the Fuzzy C-Means (FCM) algorithm and introduce a fuzzy gap statistic. We compare the gap statistic performance versus the partition coefficient and entropy indices introduced by Bezdek, the Xie-Beni and extended Xie-Beni indices, and the Fukuyama-Sugeno index. We show that the fuzzy gap statistic is more robust than the ordinary gap statistic for the IRIS data set, and we show promising results when comparing the gap statistic to the traditional fuzzy clustering indices.

KEY WORDS

Fuzzy C-Means, gap statistic, cluster validity.

1. Introduction

The Fuzzy C-Means (FCM) algorithm [1] is an unsupervised clustering algorithm, which assigns a fuzzy membership to each data point based upon its proportion of membership to each of the identified cluster means. The problem, as with many unsupervised clustering algorithms, is in determining the optimal number of clusters, which is usually a user-specified parameter. Sub-optimal clustering performance results from specifying too many or too few clusters. Many cluster validity indices have been devised to address this issue such as Bezdek's partition coefficient and entropy indices [2], the Xie-Beni and extended Xie-Beni indices [3], and the Fukuyama-Sugeno index [4]. Each of these cluster validity indices provides a unique measure of the clustering quality for a specified number of clusters. The optimal number of clusters is assessed by comparing the clustering quality measured by these indices as the number of user-specified clusters is altered.

The gap statistic, introduced by Tibshirani [5], was shown to outperform other cluster validity indices for the *K*-means algorithm, based upon identifying the a priori, known number of clusters within the dataset. The gap statistic works by computing an error measure, based upon the pooled within-cluster sum of square distances around the cluster means, for the dataset and comparing this to the expected value of the measure for a null-model (contains only a single cluster) generated from a reference distribution. Tibshirani further shows that a uniform distribution is the ideal reference distribution in terms of maximizing the gap statistic sensitivity. Like the cluster validity indices, the gap statistic is measured for several different, user-specified number of clusters. A heuristic is employed to find the optimal number of clusters as the minimum number of clusters which provides a decrease in the gap statistic. The gap statistic is applicable to any clustering algorithm for which the within-cluster sum of square distances is optimized.

The fuzzy partition matrix generated by the Fuzzy C-Means algorithm (FCM) must be made crisp prior to application of the existing gap statistic algorithm. In this paper, a fuzzy gap statistic is introduced, which does not discard the information contained within the fuzzy partition matrix while estimating the optimal number of clusters. Our experiments demonstrate some improved robustness, when comparing the fuzzy gap statistic with the ordinary gap statistic. Furthermore, our results indicate that there is merit in examining the fuzzy gap statistic more carefully, since it has produced some very promising results for the benchmark IRIS dataset, compared to the other fuzzy cluster validity indices.

In Section 2 we provide a brief review of the Fuzzy C-Means algorithm and the gap statistic along with an introduction to the fuzzy gap statistic for FCM. In Section 3 we provide a discussion of the experiments and results. The summary of our work and conclusions are included in Section 4.

2. Applying the Gap Statistic to FCM

2.1 Fuzzy C-Means

The Fuzzy C-Means (FCM) algorithm is an unsupervised clustering algorithm, which partitions data into clusters using fuzzy membership. Each data point, $\mathbf{x}_j$, is assigned a fuzzy membership, based upon its membership to each cluster, $u_{jk} = \left\{ u_{jk} \in \mathbb{R} \mid 0 \le u_{jk} \le 1, \sum_{k=1}^{c} u_{jk} = 1.0 \right\}$ where, c, is the number of clusters, j, is the data point index, k, is the cluster index, and $u_{jk} \in \mathbf{U}$ is the membership value of data-point j to cluster k where $\mathbf{U}$ is the fuzzy partition matrix. Each cluster, k, is also assigned a representative center, $\mathbf{v}_k \in \mathbb{R}^n$. By employing fuzzy membership, the FCM algorithm allows partial membership of each data point to each cluster. The goal of FCM is to create a partitioning of the data, for a specified number of clusters, that minimizes the following objective function

$$J_m = \sum_{k=1}^{n} \sum_{i=1}^{c} \left(u_{ik} \right)^m \left\| \mathbf{x}_k - \mathbf{v}_i \right\|_A^2 \qquad (1)$$

where, A, refers to a generalized norm (e.g., $A = 2$, refers to the Euclidean norm), and, m, is the fuzzy factor index and controls the shape of the membership functions. As $m \to 1$, the partitions become crisp and FCM approaches the behaviour of the K-means algorithm.

Several indices exist to determine the overall “goodness of fit” for a specified number of clusters. These indices include Bezdek’s partition coefficient, v_{PC}, and partition entropy, v_{PE}, indices; the Xie-Beni index, v_{XB}, and extended Xie-Beni index $v_{XB,m}^{FCM}$; and the Fukuyama-Sugeno index, v_{FS}. Note that the listed indices all utilize the information contained in the fuzzy partition matrix, highlighting the possible importance of including this information in a gap statistic implementation.

2.2 Gap Statistic

The gap statistic is a cluster validity measure based upon a statistical hypothesis test. The null hypothesis states there is a single cluster within the data. The gap statistic works by comparing the within-cluster similarity for the data set under consideration versus the expected within-cluster similarity of data points generated from a null-model reference distribution at each value of k.

An error measure, or within-cluster similarity, W_k, is defined as

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,i' \in C_r} d_{ii'} , \qquad (2)$$

where W_k is the pooled with-in cluster sum of squares around the cluster means, k, is the number of clusters, C_r, is the set of points in cluster r, $d_{ii'}$ is the Euclidean distance between points i and i', and n_r is the number of points within cluster r. The gap statistic is, then, defined as

$$Gap(k) = E\left\{ \log\left(W_k^* \right) \right\} - \log\left(W_k \right) \qquad (3)$$

where $E\left\{ \log\left(W_k^* \right) \right\}$ is the expected value of the error measure, W_k, for a null-model reference distribution. Equation (3) is then used to generate the gap curve. The error measure, W_k, will decrease monotonically as the number of clusters, k, is increased; however, it will decrease more rapidly when the optimal number of clusters is reached and, then, continue to decrease at its previous rate. The gap curve, representing the difference between the expected error measured and error measure, will increase sharply just prior to reaching the optimal number of clusters and, then, level out. The goal of the gap statistic heuristic is to discover this point, which is often referred to as the “elbow” in the gap curve. Figure 8, referred to later in the experimental results, depicts a typical example of the gap curve.

Samples originating from the null-model reference distribution are generated by randomly sampling a uniform distribution either covering the extent of each dimension of the dataset under test or the principal components of the dataset to create a more compact distribution.

To obtain $E\left\{ \log\left(W_k^* \right) \right\}$, a Monte Carlo simulation is performed B times where W_k^* is measured from a new sample of the reference distribution. The expected value is, then,

$$E\left\{ \log\left(W_k^* \right) \right\} = \frac{1}{B} \sum_{b=1}^{B} \log\left(W_{kb} \right). \qquad (4)$$

The simulation error associated with the Monte Carlo simulation can be computed as

$$s_k = sd\left(k\right) \sqrt{1 + 1/B} \qquad (5)$$

where the standard deviation is

$$sd\left(k\right) = \left[\frac{1}{B} \sum_b \left(\log\left(W_{kb} \right) - \overline{\log\left(W_{kb} \right)} \right)^2 \right]^{\frac{1}{2}}. \qquad (6)$$

Once the gap statistic is computed for each number of clusters, k, the optimal number of clusters, $\hat{k}$, is the smallest, k, such that

$$Gap\left(k\right) \geq Gap\left(k+1\right) - s_{k+1}. \quad (7)$$

2.3 Fuzzy Gap Statistic

An obvious solution for applying the gap statistic to FCM is to form a crisp partition by “hardening” the fuzzy partition matrix prior to computing the gap statistic. However, this methodology discards potential information contained in the fuzzy memberships. As stated in [6], fuzzy sets “provide information about overlap and substructure within the data.” Detection of this substructure may be important when determining the optimal number of clusters.

The goal, here, is to create an error measure for use with the gap statistic that incorporates the fuzzy membership data. Equation (2) can be modified to an equivalent form where the summation of distances from each point to its corresponding cluster center is computed

$$\begin{aligned}\sum_{i=1}^{n_k}\left\|x_i - v_k\right\|_2^2 &= \sum_{i=1}^{n_k}\left\langle x_i - v_k, x_i - v_k\right\rangle \\ &= \sum_{i=1}^{n_k}\left\langle x_i - \frac{1}{n_k}\sum_{j=1}^{n_k} x_j, x_i - \frac{1}{n_k}\sum_{j=1}^{n_k} x_j\right\rangle\end{aligned} \quad (8)$$

where $\left\langle , \right\rangle$ denote the usual vector inner product. Through some additional manipulation, we obtain

$$\begin{aligned}&\sum_{i=1}^{n_k}\left\langle x_i, x_i\right\rangle - \frac{1}{n_k}\left[\sum_{i=1}^{n_k}\left\langle x_i, x_i\right\rangle + 2\sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k}\left\langle x_i, x_j\right\rangle\right] \\ &= \frac{n_k - 1}{n_k}\sum_{i=1}^{n_k}\left\langle x_i, x_i\right\rangle - \frac{2}{n}\left[\sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k}\left\langle x_i, x_j\right\rangle\right]\end{aligned} \quad (9)$$

On the other hand, the sum of all pair-wise distances can be written as

$$\begin{aligned}&\sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \\ &= \sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k}\left\langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j\right\rangle \\ &= \sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k}\left\langle \mathbf{x}_i, \mathbf{x}_i\right\rangle - 2\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle + \left\langle \mathbf{x}_j, \mathbf{x}_j\right\rangle \\ &= \left(n_k - 1\right)\sum_{i=1}^{n_k-1}\left\langle \mathbf{x}_i, \mathbf{x}_i\right\rangle - 2\sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k}\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle\end{aligned} \quad (10)$$

As a result, we obtain

$$\sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 = n_k\sum_{i=1}^{n_k}\left\|\mathbf{x}_i - \mathbf{v}_k\right\|_2^2 \quad (11)$$

which indicates that the sum of square of pair-wise distances within a particular cluster is equal to the sum of square of distances to the cluster mean times the number of data points within the cluster, n_k.

As a result, we can create a new formulation for the error measure, W_k, as

$$W_k = \sum_{r=1}^{k}\sum_{i\in C_r}\left\|x_i - v_r\right\|_2^2 \quad (12)$$

where C_r is the set of data points belonging to cluster r, and v_r is the cluster center. Comparing (12) with (1), we observe that J_m is similar to W_k, and, in fact, is referred to as the pooled within cluster sum of square distances within the FCM literature. This suggests J_m can be used as an error measure to create a fuzzy gap statistic.

$$J_{m,k} = \sum_{r=1}^{k}\sum_{i=1}^{n}\left(u_{ri}\right)^m\left\|\mathbf{x}_i - \mathbf{v}_r\right\|_A^2 \quad (13)$$

where, m, is the fuzzy factor index, k, is the number of clusters, n, is the number of data points, $\mathbf{x}_i$, is the ith data point, and $\mathbf{v}_r$ is the cluster center for cluster r. As an additional motivation, note that as $m \to 1$, $J_{m,k}$ approaches the behavior of W_k. The gap statistic is now defined as

$$Gap(k) = E\left\{\log\left(J_{m,k}^*\right)\right\} - \log\left(J_{m,k}\right) \quad (14)$$

where $E\left\{\log\left(J_{m,k}^*\right)\right\}$ is the expected value of the error measure, $J_{m,k}$, for a null-model reference distribution

$$E\left\{\log\left(J_{m,k}^*\right)\right\} \cong \frac{1}{B}\sum_{b=1}^{B}\log\left(J_{m,k,b}\right). \quad (15)$$

The standard deviation now becomes

$$sd\left(k\right) = \left[\frac{1}{B}\sum_{b}\left(\log\left(J_{m,k,b}\right) - \overline{\log\left(J_{m,k,b}\right)}\right)^2\right]^{\frac{1}{2}}. \quad (16)$$

The heuristic for selecting the optimal number of clusters based upon the gap statistic remains the same as before.

3. Experiments and Results

The IRIS data set from the UCI repository [7] and the 4-NORM data set from Bezdek [2] are used for comparison of the performance of the cluster validity indices versus the gap statistic and are commonly seen in the literature for this type of comparison. The 4-NORM data set consists of 4 clusters of 4-variate data ($x \in \mathbb{R}^4$) with the center of each cluster projected 3 units along one of each of 4 axes. The covariance of each cluster is $\Sigma^{-1} = \mathrm{I},$ and there are 200 data points in each cluster for a total of 800 data points.

With these experiments, and for the same dataset, it is recognized that FCM will produce different results based upon its random initialization. Furthermore, the Monte Carlo simulations performed as part of the gap statistic measure will also yield different results. Therefore, we performed 50 trials on each data set for each of the cluster validity mechanisms considered.

For each case, we vary the fuzzy factor index, $m,$ over the values of 1.2, 2.0 and 7.0. While the ideal range suggested by Pal [2] is [1.5, 2.5], we are interested in examining the robustness of fuzzy gap statistic versus the gap statistic over a wider range of values.

For the IRIS data set and $m = 2.0,$ and $m = 1.2,$ both the gap statistic and fuzzy gap statistic perform well, and, in fact, both methods consistently predict 3 clusters. Note that, with the exception of the Fukuyama-Sugeno index, v_{FS} , the remaining cluster validity indices consistently identify 2 clusters.

For the case $m = 7.0,$ for the IRIS dataset, the fuzzy gap statistic consistently identifies 3 clusters; however, the ordinary gap statistic identifies both 3 and 4 clusters with more samples at 4 clusters, suggesting that the fuzzy gap statistic is more robust as the fuzzy factor, $m,$ is varied. The remaining cluster validity indices, with the exception of the extended Xie-Beni index, $v_{XB,m}^{FCM},$ tend to consistently identify 2 clusters.

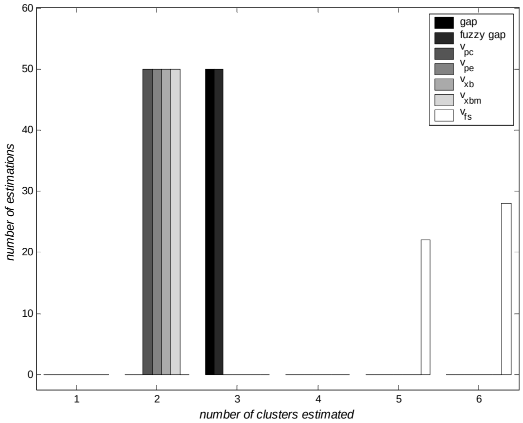

Figure 1 Histogram of the number of clusters identified by each index out of 50 trials for the IRIS data set and m = 1.2. Note that a value of 6 implies cluster sizes >= 6 and the order (top to bottom) in the legend represents the order (left to right) within in each group of the bar plot.

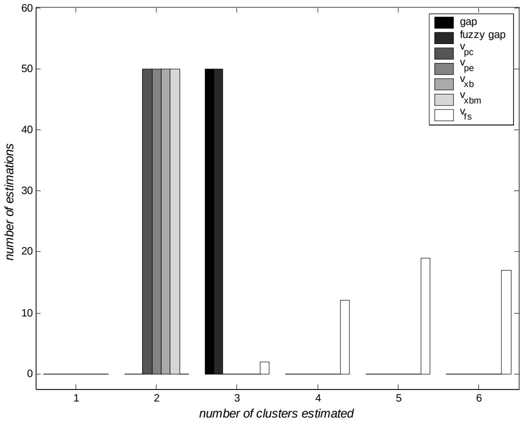

Figure 2 Histogram of the number of clusters identified by each index out of 50 trials for the IRIS data set and m = 2.0. Note that a value of 6 implies cluster sizes >= 6 and the order (top to bottom) in the legend represents the order (left to right) within in each group of the bar plot.

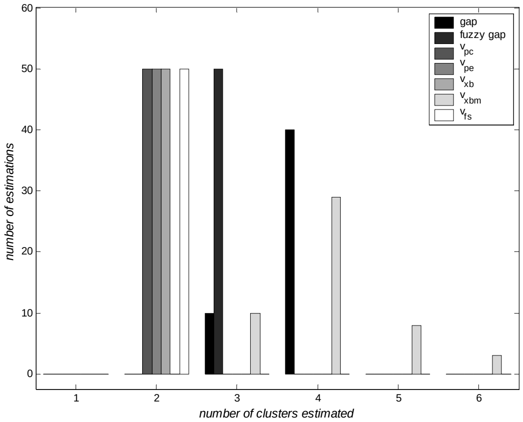

Figure 3 Histogram of the number of clusters identified by each index out of 50 trials for the IRIS data set and m = 7.0. Note that a value of 6 implies cluster sizes >= 6 and the order (top to bottom) in the legend represents the order (left to right) within in each group of the bar plot.

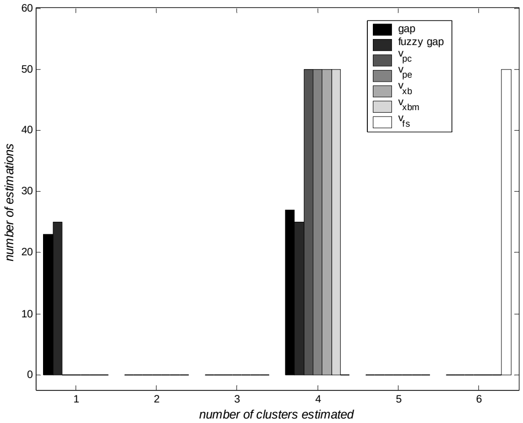

Figure 4 Histogram of the number of clusters identified by each index out of 50 trials for the 4-NORM data set and m = 1.2. Note that a value of 6 implies cluster sizes >= 6 and the order (top to bottom) in the legend represents the order (left to right) within in each group of the bar plot.

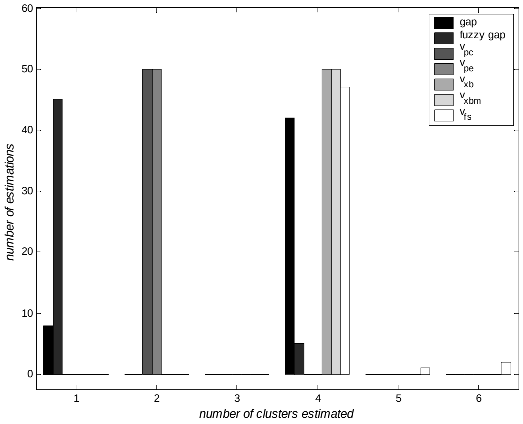

Figure 5 Histogram of the number of clusters identified by each index out of 50 trials for the 4-NORM data set and m = 2.0. Note that a value of 6 implies cluster sizes >= 6 and the order (top to bottom) in the legend represents the order (left to right) within in each group of the bar plot.

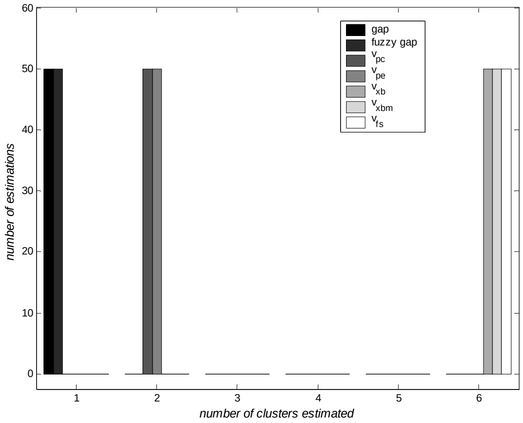

Figure 6 Histogram of the number of clusters identified by each index out of 50 trials for the 4-NORM data set and m = 7.0. Note that a value of 6 implies cluster sizes >= 6 and the order (top to bottom) in the legend represents the order (left to right) within in each group of the bar plot.

In summary, for the IRIS data set, both the gap statistic and fuzzy gap statistic are able to identify 3 clusters within the IRIS data set while the remaining indices appear to report 2 clusters or are inconsistent. As m is increased, the fuzzy gap statistic appears to be the most robust since the ordinary gap statistic performance begins to break down at $m = 7.0$ in terms of consistency.

The results appear quite different for the 4-NORM data set (Figures 4 through 6). First, we note that all of the FCM cluster validity indices, with the exception of v_{FS}, identify 4 clusters at $m = 1.2$, are split between identifying 2 and 4 clusters at $m = 2.0$, and are split between identifying 2 and 6 clusters at $m = 7.0$. Note that the FCM cluster validity indices are incapable of identifying a single cluster and this is an advantage of the gap statistic.

The gap statistic, on the other hand, performs quite differently. Overall, we note that both the gap statistic and fuzzy gap statistic choose between either 1 or 4 clusters at $m = 1.2$ and $m = 2.0$. The fuzzy gap statistic also appears to choose 1 cluster more often than the gap statistic for the case $m = 2.0$. For the case, $m = 7.0$, both forms of the gap statistic choose 1 cluster. It would also appear that the fuzzy gap statistic tends towards the decision to report 1 cluster quicker than the ordinary gap statistic as m is increased.

It is interesting, however, to observe the gap curve for the 4-NORM dataset, as depicted in Figures 7 through 8, for the fuzzy gap statistic. Tibshirani reports that the heuristic (7) used by the gap statistic is not designed to handle the case where smaller sub-clusters are contained within a larger cluster, and, in these cases, the gap curve, generated by (3), should be manually reviewed. Note that there is a distinct "elbow" at $k = 4$. We also see that there is a "flat" region, within the simulation error, when progressing from $k = 1$ to $k = 2$. This gap curve can be interpreted to indicate there are 4 sub-clusters within one larger cluster, which is arguably correct for the 4-NORM dataset. Furthermore, the gap curve provides a stronger indication for 4 clusters than a single cluster, which the heuristic (7) fails to recognize since it stops as soon as the single cluster is recognized. This suggests additional research is needed to investigate possible modifications to (7) in these cases. In summary, the gap curve indicates the gap statistic is correctly identifying the number of clusters for the 4-Norm dataset despite the failure of the heuristic (7).

At $m = 7.0$, it appears that all FCM cluster validity indices are no longer able to distinguish the 4 clusters within the data and become inconsistent as they are unable to report a single cluster, while both forms of the gap statistic identify a single cluster. Observing the gap curve in Figure 9, we see that there is no longer any evidence for clusters within the data, supporting the claim of a single cluster, in this case.

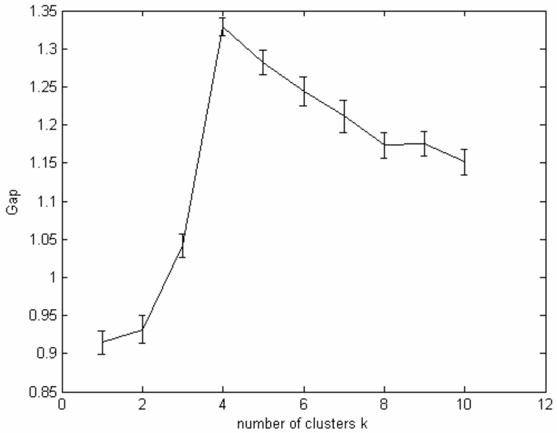

Figure 7 The fuzzy gap curve when applied to the 4-NORM data set when m = 1.2. The "elbow" in the curve at k = 4 is readily evident as well as a flat response at k = 1 indicating 4 sub-clusters within one larger cluster.

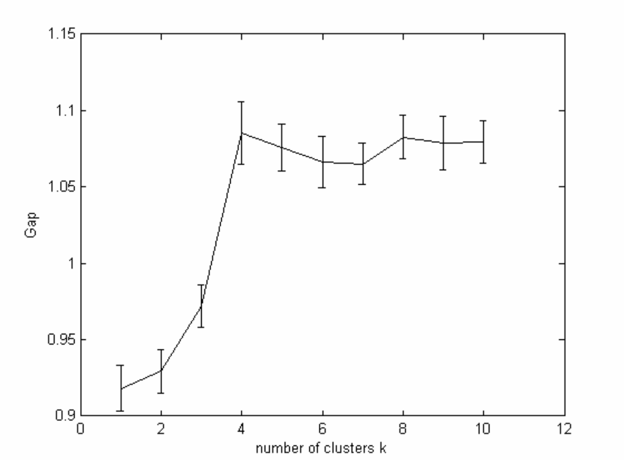

Figure 8 The fuzzy gap curve when applied to the 4-NORM data set when m = 2.0. The "elbow" in the curve at k = 4 is readily evident as well as a flat response at k = 1 indicating 4 sub-clusters within one larger cluster.

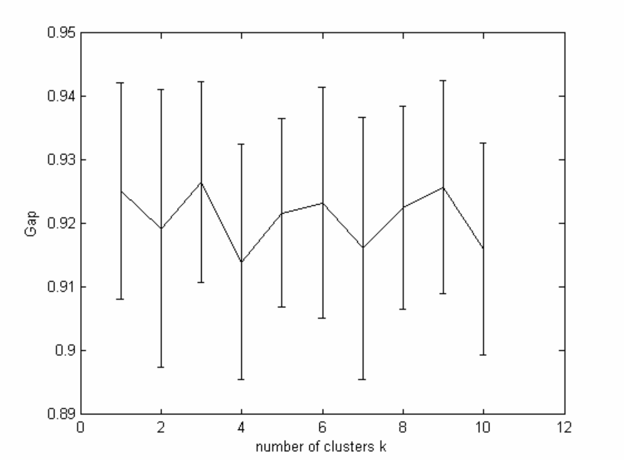

Figure 9 The fuzzy gap curve when applied to the 4-NORM data set when m = 7.0. The flat response, here, indicates a null-model or single cluster.

4. Conclusion

Overall, we observe promising results for the fuzzy gap statistic that are worth investigating further. Predominantly, we note improved robustness as the fuzzy factor, *m*, is changed between 1.2 and 7.0 for the IRIS data set. Although the appropriate number of clusters is still a rather subjective issue, it is interesting to note that only the gap statistic identifies 3 clusters for the IRIS data set. It seems, then, that future research of this concept may be merited with more focused investigation on the specific impacts of including a fuzzy factor in the gap statistic in terms of sensitivity as well as into alternative heuristics for selecting the optimal number of clusters using the gap statistic.

Acknowledgment

Chris Sentelle gratefully acknowledges the support of the College of Engineering & Computer Science and the I2Lab at the University of Central Florida. This work was supported in part by the NSF grants: CRCD: 0203446, CCLI: 0341601, DUE: 05254209, IIS-REU: 0647120, and IIS-REU: 0647018.

References

[1] J. C. Bezdek, Numerical taxonomy with fuzzy sets, *Journal of Mathematical Biology, 1*(1), 1974, 57-71.

[2] N. R. Pal & J. C. Bezdek, On cluster validity for the fuzzy c-means model, *IEEE Trans. Fuzzy Systems*, *3*(3), 1995, 370-379.

[3] X. L. Xie & G. A. Beni, Validity measure for fuzzy clustering, *IEEE Trans. Pattern Analysis Machine Intell.*, *13*(8), 1991, 841-847.

[4] Y. Fukuyama & M. Sugeno, A new method of choosing the number of clusters for the Fuzzy C-Means method, *Proc. 5th Fuzzy Syst. Symp*., 1989, 247-250.

[5] R. Tibshirani, G. Walther, & T. Hastie, Estimating the number of clusters in a dataset via the gap statistic, *J. R. Statist. Soc. B*, *63*(2), 2001, 411-423.

[6] H. Hassar & A. Bensaid, Validation of fuzzy and crisp c-partitions. *Conf. of North-American Fuzzy Information Processing Society (NAFIPS)*, New York, NY, 1999, 342-346.

[7] D. J. Newman, S. Hettich, C. L. Blake, & C. J. Merz, 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, 1998.

[8] J. T. Tou & R. C. Gonzalez, *Pattern Recognition Principles* (Reading, MA: Addison-Wesley, 1974).

[9] C. Arima, K. Hakamada, M. Okamoto, T. Hanai, Validity index for Fuzzy K-Means clustering using the gap statistic method, *Sixteenth International Conference on Genome Informatics*, Pacifico Yokohama, Japan, 2005.

[10] J. C. Bezdek, J. M. Keller, R. Krishnapuram, L. I. Kuncheva, N. R. Pal, Will the real Iris data please stand up?, *IEEE Trans. Fuzzy Systems*, *7*(3), 1999, 368-369.

[11] T. Hastie, R. Tibshirani, & J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York, NY: Springer-Verlag, 2001).

[12] G. W. Milligan & M. C. Cooper, An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, *50*, 1985, 159-179.

[13] _______, *Experiments with K-means, Fuzzy C-Means and Approaches to Choose K and C* (Orlando, FL: Main Library, Orlando University Archives, 2006).