Investigation of Internal Validity Measures for K-Means Clustering

Jonathan Baarsch and M. Emre Celebi

***Abstract*—Clustering is a fundamental task in data mining and knowledge discovery. The most widely used technique for clustering is the k-means algorithm, which is dependent on the choice number of clusters, k. In unsupervised situations, the choice of an appropriate value for k is difficult. To overcome this challenge, validity measures attempt to determine how accurately the clusters reflect the data. However, numerous validity measures proliferate, and different measures often produce disparate results. This paper reports an experiment to evaluate commonly used cluster validity measures, including Dunn, Davies-Bouldin, Calinski-Harabasz, Silhouette, Point Bi-serial, PBM, and Sum-of-Squares. These measures were applied to k-means clusterings of 125 artificially generated data sets. The Sum-of-Squares method was found to be the most effective for predicting an appropriate value for k. Silhouette was found to be a good alternative, and Calinski-Harabasz and Davies-Bouldin both made only moderate showings compared to the other two. Dunn, Point Bi-serial, and PBM performed quite poorly. The results also suggest that validity measures could be used as explanatory tools in their own right.**

***Index Terms*—cluster analysis, cluster validation, k-means**

I. Introduction

It is beyond cliché to mention the current importance of clustering to data mining and knowledge discovery [1]. In our data-saturated world, and particularly in the areas of computer and information sciences, grouping multi-dimensional data into clusters for classification or more efficient processing is ubiquitous. The most widely used technique for this clustering is the k-means algorithm [1][2]. The k-means algorithm is efficient and effective, but it suffers from some frequently lamented shortcomings. Specifically, clustering by k-means favors hyper-spherical clusters, since the algorithm typically uses some variation on Euclidean distance from the cluster center as its primary clustering criteria. Often, however, hyper-spherical results are to be desired. A more persistent problem with k-means that occurs even when the actual structure of the data is hyper-spherical, is that the outcome of the clustering algorithm is dependent on certain parameters, most significantly for the purpose of this paper, on the number of clusters, or "k".

The first step in the k-means algorithm is to choose the number of clusters into which the data will be divided. The initial choice of k is largely an interpretive decision. Successive runs of k-means can optimize the division of the data for any given number of clusters, but choosing an initial k-value presumes prior knowledge of the data's structure. Without such knowledge, it is difficult to know the proper number of clusters—but *with* such knowledge, the need for the clustering is diminished. It is even conceivable that certain data sets might have more than one "natural" clustering. With simple data, i.e. 2- or perhaps even 3-dimensional, it is possible for the human eye to pick out possible candidates for k, but as the dimensionality of the data increases, it becomes increasingly difficult to determine a proper value for k. And, again, in the cases where the eye can pick them out, there isn't a need for a clustering algorithm in the first place.

It is for this reason that considerable attention has been given over the past 40 years to the subject of cluster validation—a process which attempts to evaluate a particular division of data into clusters. While many sources claim there are three types of cluster validation: internal, external, and relative [3], there are really only two that have received attention from the research community: *external validation*, which measures the clusters against some pre-determined structure of the data, and *relative validation*, which measures various attempts at clustering a set of data against each other, usually by way of a "score." External validation presumes a pre-known structure of the data and so is not particularly relevant to this discussion. Relative validation, in contrast, is typically used to solve the problem of choosing a good k value. To the extent that validation can successfully choose, from a group of clustering attempts with various k, the k-value that best models the actual structure of the data, cluster validation is an integral step in the process of clustering data in unsupervised situations.

Because the k-means algorithm is often applied in order to interpret and understand data about which little is known (unsupervised learning), validation becomes a *de-facto* last step in the clustering process. First, the data is clustered into various clusters of different k, and then validation methods select the "best" clustering from those. Therefore, validation is crucial to a successful division of data when the number of clusters is unknown.

This project aims to implement some of the most popular and successful validation methods in order to compare the validations against each other. There have been dozens of validation measures proposed over the past 40 years, and even though research in this area is still going strong, with new methods proposed every year, some of the earliest algorithms have been shown to be the most effective [4][5]. This project is modeled after a few independent studies that have

Manuscript received November 22, 2011; revised January 10, 2012. This publication was made possible by grants from the Louisiana Board of Regents (LEQSF2008-11-RD-A-12) and US National Science Foundation (0959583, 1117457).

The authors are with the Department of Computer Science, Louisiana State University, Shreveport, LA, USA. E-mail: {jbaarsch, ecelebi}@ lsus.edu.

measured validation techniques against each other [4][5]. Although most papers that propose new clustering validation methods also measure their algorithms against other validation methods, (for example, [6]), independent tests, which have no agenda of their own, are often more reliable.

II. Description of Validation Measures

Of the dozens of validation measures, this project had to choose only a few. Choices were made on the basis of three criteria: success of the measure in the literature, popularity of the measure, and simplicity/efficiency of implementation. The algorithms should be efficient enough to be processed quickly even on high-dimensional and large data sets, but also simple enough for quick implementation.

The two nearly universal general criteria used by all validity measures to evaluate clusters are compactness and separation. A good clustering will create clusters with points that are similar or "close" in quantifiable Euclidean terms to one another (compact), but different or "distant" from points in other clusters (separation). Almost all of the most common and successful validity measures try to measure compactness and separation and relate them to one another, either maximizing the inter-cluster distance, or minimizing the intra-cluster distance, or maximizing/minimizing a ratio between measures of both qualities. While the methods to determine separation or compactness can sometimes vary, it should be noted that most validity measures, like the k-means clustering algorithm itself, favor hyper-spherical clusters, using the cluster centers as the basis for measuring compactness, if not also separation.

The methods employed in this project were those commonly known as Dunn [7], Davies-Bouldin [8], Silhouette [9], Calinski-Harabasz [10], Sum-of-Squares [6], Point Biserial [11], and PBM [12]. The following subsections provide descriptions of these validity measures.

A. Dunn

Perhaps the most frequently cited measure, the Dunn method provides a score based on the square root of the minimum distance between any two clusters (measuring separation) divided by the square root of the maximum distance between any two points in the same cluster (measuring compactness). The distance between two clusters is measured by the distance between the two closest points, thus the ratio becomes:

$$\sqrt{\text{Min Intercluster Dist}}/\sqrt{\text{Max Intracluster Dist}}$$

The higher the value, the "better" the clustering will be. Because the measure only uses minimum and maximum values rather than averages or aggregates, the Dunn method is highly susceptible to influence from noise, outliers, or two clusters that happen to be close together.

B. Davies-Bouldin

The Davies-Bouldin (DB) method also relates compactness to separation. DB compares each cluster to every other cluster based on a function measuring similarity in which for each pair of clusters, the sum of the average distances of each point in the two clusters to its respective center (that is, a sum of the dispersion of the two clusters, measuring compactness) is divided by the distance between the two cluster centers (measuring separation). The maximum values of this function for each cluster are averaged, resulting in a score:

$$\frac{1}{k}\sum_{i=1}^{k} R_i$$

where $R_i = \max R_{ij},\ i \neq j.\ R_{ij} = (S_i + S_j)/M_{ij}.\ S_i$ is the sum of the average distances from each point in cluster i to the centroid of its cluster, and M_{ij} is the distance between the two cluster centers.

For DB, a lower score will be the result of less dispersion within clusters and more distance between clusters. Unlike Dunn, DB uses cluster centroids to represent clusters in order to measure separation. Because the score uses the maximum comparison for each cluster, the measure is built upon "worst-case" situations. DB divides compactness by separation, meaning that as clusters become more compact and more separated, the DB value will shrink.

C. Silhouette

The Silhouette method also relates compactness to separation, but unlike DB, Silhouette is based on the mean score for every point in the data set. Each point's individual score is based on the difference between the average distance between that point and every other point in its cluster and the minimum average distance between that point and the other points in each other cluster. This difference is then divided by a normalizing term, which is the greater of the two averages:

$$\frac{1}{N}\sum_{i=0}^{N} s_{x_i}$$

where N is the number of points in the data set and:

$$s_{x_i} = (b_{q,i} - a_{p,i})/max\{a_{p,i}, b_{p,i}\}.$$

If x_i is a point in cluster p, then,

$$b_{q,\,i} = \min d_{q,i} \ \text{ where } d_{q,i}$$

is the average distance between point x_i and every point of cluster q. On the other hand, $a_{p,i}$ is the average distance between point x_i and every other point of cluster p. Unlike Dunn and DB, the Silhouette measure relates separation to compactness by subtraction rather than division. As clustering improves, the score will approach 1.

D. Calinski-Harabasz

The Calinski-Harabasz (CH) method has been one of the most successful in independent studies. The method is based on a relationship between a "between cluster scatter matrix" (BCSM) and a "within cluster scatter matrix" (WCSM):

$$\frac{trace(BCSM)}{trace(WCSM)} * \frac{N-k}{k-1}$$

The trace BCSM is merely the sum of the squares of the distances between the center of each cluster and the centroid of the data set, weighted by the size of the cluster. The trace

WCSM is the sum of the squares of the distances between the center of each cluster and every point in the cluster.

Like DB, CH uses cluster centers for calculating separation; however, separation is measured according to the center of the data set, rather than particular clusters. The normalization factor, $(N-k)/(k-1)$, will diminish the score as k increases.

E. Sum-of-Squares

The sum-of-squares (SS) method is a simple adaptation of the CH method:

$$\frac{trace(WCSM)}{trace(BCSM)} * k$$

Thus, the measure reverses the relationship between separation and compactness, and the normalization factor changes much more dramatically as k increases. Since SS divides compactness by separation, like DB, a lower score indicates better clusterings.

F. Point Bi-serial

The Point Bi-serial (PB) method finds the difference between the average intra-cluster distance (that is, the average distance between each point in the cluster, a measure of compactness) and the average inter-cluster distance (that is, the distance between each point and all the other points in the data set that are not in that point's cluster, a measure of separation):

$$\left(\overline{d_s} - \overline{d_c}\right) * \frac{\sqrt{(\alpha * \beta)/x^2}}{\sigma}$$

where $d_c$is the distance between each point and every other point in its respective cluster and d_s is the distance between each point and every other point not in its cluster. In the normalization figure, α is the number of intra-cluster distances, and β is the number of inter-cluster distances and *x* is the number of point pairs in the data set. σ is the standard deviation of the distances. This measure is like Silhouette, except that it measures separation from all non-cluster sharing points, rather than only those of the closest cluster.

G. PBM

The PBM score relates a figure of compactness, measured as the sum of the distances between each point and its cluster centroid, to a measure of separation, calculated as the maximum distance between any two cluster centers, normalized over a measure of dispersion, calculated as the sum of the distances between all points:

$$\left(\frac{1}{k} * \frac{\alpha}{\beta} * \gamma\right)^2$$

where α is the sum of the distances between each point and its cluster center, β is the sum of the distances between each point and the centroid of the data set, and γ is the maximum distance between any two cluster centers. Separation is measured by the greatest distance between any two clusters, so this measure favors clusterings that have at least two well-separated clusters.

III. Experimental Results

A cluster validity measure can be tested on two types on data sets: real-world or artificial. Even though real-world data sets are useful insofar as they represent the complexity of the real world, the class attributions of such data sets are not generally based on a structural analysis of the data itself. As a consequence, the correspondence between optimal clustering and the nominally correct clustering is questionable. Therefore, we decided to use artificial data sets where the structure of the data could be controlled and known a priori.

In order to compare the seven validity measures, we artificially generated 125 data sets using the Clustering Algorithms' Referee Package [13]. The size of the data sets ranged from 256 to 4096 points: 256, 512, 1024, 2048, and 4096. The number of attributes ranged from 2 to 32: 2, 4, 8, 16, and 32. Finally, the number of clusters ranged from 2 to 10: 2, 4, 6, 8, and 10.

The data was clustered using the k-means algorithm [14]. The algorithm starts with k initial centers, typically chosen uniformly at random from the data points. Each point is then assigned to the nearest center, and each center is recalculated as the mean of all points assigned to it. These two steps are repeated until a change of .1% or less in the sum-of-squared error is observed or 20 iterations, whichever occurred first. As for the selection of the initial centers, we employed the farthest-first method [15] in which the first center is chosen arbitrarily and each of the remaining centers is chosen as the point that is the farthest from the already chosen centers.

For each data set, k-means was run successively to create optimal clusterings of the data for k values between 2 and 15. These optimal clusterings were tested by the seven validity measures, resulting in a set of 14 "scores" for each validity measure, one each for k = 2 through 15. These scores were then compared against each other to find the "best" k value for the clustering according to the validity measure in question. Finally, the result was compared to the "correct" number of clusters. If the validity measure correctly predicted the number of clusters, then it scored a "correct" tally. If not, the absolute difference between the prediction and the "true" number of clusters was calculated. After processing all 125 data sets, these two tallies, the number of correct predictions and the running sum of the deviation were recorded, with the deviation being divided by the number of data sets (125) to provide an average absolute deviation. Table 1 gives the results. Here, the last three columns, from left to right, indicate the number (percentage) of correct predictions, average (standard deviation) of the absolute differences, and average of the absolute differences for only the mispredictions, respectively. It can be seen that SS is the best predictor, which is followed closely by Silhouette.

TABLE I
COMPARISON OF THE INTERNAL VALIDITY MEASURES

Dunn	46 (36.8%)	2.272 (3.024)	3.595
DB	47 (37.6%)	1.456 (1.657)	2.333
Silhouette	64 (51.2%)	0.968 (1.563)	1.984
CH	54 (43.2%)	2.184 (2.735)	3.845
SS	**77 (61.6%)**	**0.808 (1.577)**	**2.104**
PB	19 (15.2%)	3.600 (3.446)	4.245
PBM	36 (28.8%)	3.024 (2.768)	4.247

IV. DISCUSSION

When compared to the results of other comparative studies, one of the most striking results is the lack of accuracy of the results. Although SS could predict the correct number of clusters more than half the time, most of the others could not, even SS was wrong nearly 40% of the time.

There are three reasons we believe the scores are as low as they are. First, when investigating the results, we found irregularities in the data sets, perhaps caused by the overlap feature of the data set generation software. Additionally, we discovered that often k-means was not finding the natural clusters in the data set. Finally, some of the discrepancies and "incorrect" predictions could be the result of different interpretations of the data sets themselves.

A. Data Irregularities

Because most of the data sets were high-dimensional, and thus difficult to visualize, the analysis of the results focused on the subset of the data which was two dimensional. Fig. 1 shows such a data set with six classes. It can be seen that there is some overlap between clusters, which had two consequences. First, the clusters are close together and so the separation between them will be low. Second, it might be difficult for both clustering and validation to "split" clusters that are close together (see Figs. 2 and 3).

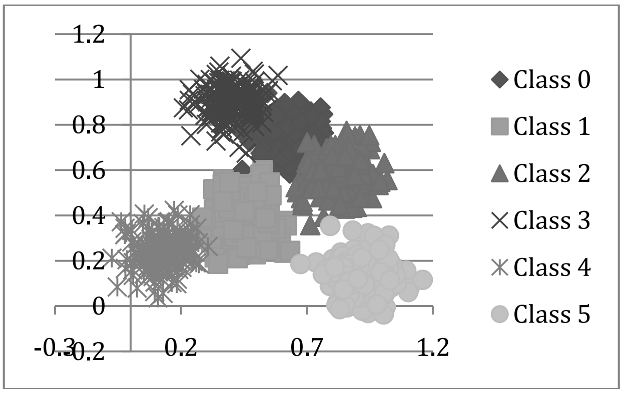

Fig. 1. 1024 2-D points divided into 6 classes

A more pernicious problem appeared when plotting the data set containing 4096 points in ten clusters. When that data set was plotted, one of the classes completely overlapped another, such that it is impossible for the eye to distinguish them without visual aid (see Fig. 4).

This phenomenon was observed in at least one of the other data sets which was divided into eight classes. Such extreme overlap could be more likely in data sets with large k and a break-down of the performance of the clustering algorithms demonstrated that all the measures showed marked improvement when the number of clusters was decreased (see Table 2). There could be other explanations for this phenomenon, but it is clear that if the data set contains clusters that completely overlap one another, neither k-means nor the validity measure will be able to distinguish them.

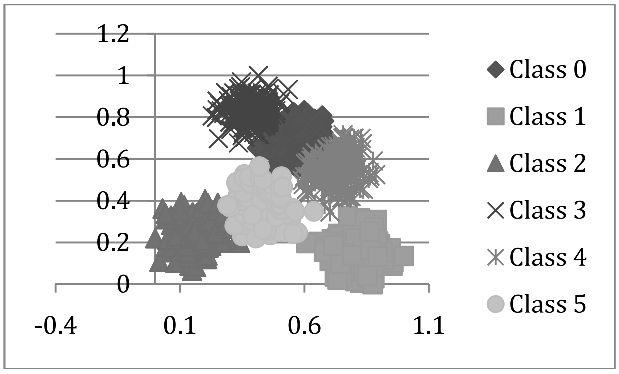

Fig 2. Data set in Fig. 3 divided into k=6 clusters

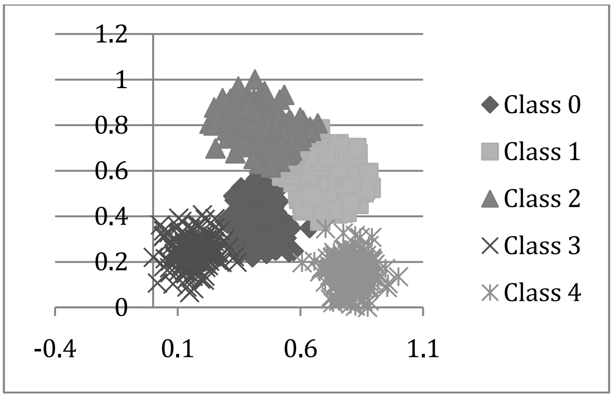

Fig. 3. Data set in Fig. 3 divided into k=5 clusters

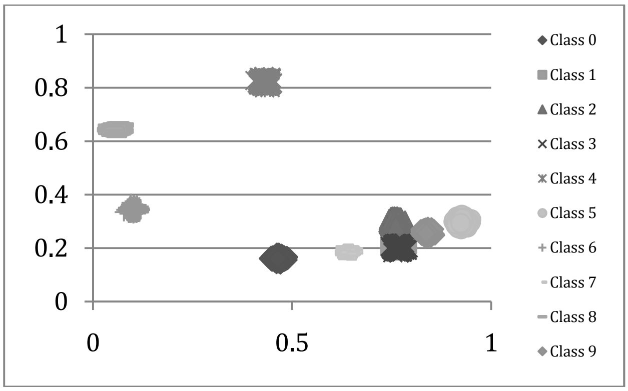

Fig. 4. 4096 2-D points divided into 10 classes

TABLE 1
ACCURACY (%) AS THE NUMBER OF CLUSTERS (K) INCREASES

	k=2	k=4	k=6	k=8	k=10
Dunn	88	44	20	20	12
DB	100	36	24	20	8
Silhouette	100	44	32	56	24
CH	100	48	32	24	12
SS	96	81	64	40	24
PB	20	4	20	16	16
PBM	100	16	16	8	4

B. *K-Means Clustering Irregularities*

Irregularities in the data sets, however, do not tell the whole story. To make matters more muddled, the clusterings created by the k-means algorithm often do not best reflect the "natural" structure of the data for any given k. Looking at the 10 cluster data set above, for instance, the eye easily makes out nine distinct clusters of points. On this data set Dunn, DB, Silhouette, CH, SS, PB, and PBM predicted 3, 4, 4, 13, 13, 10, and 13 clusters, respectively. The two reoccurring choices are 4 and 13 clusters (see Figs. 5 and 6). The four cluster division seems intuitively justifiable to the eye. The 13 cluster division, however, hardly seems as natural as the nine. However, when the clustering algorithm divided the data set into nine clusters, the result was hardly any more natural (see Fig. 7). Part of the reason for this somewhat counter-intuitive division is using the farthest-first initialization method. Fig. 8 shows how k-means clusters the same data set with random initialization, but run 100 times, choosing the clustering with the best (lowest) SSE.

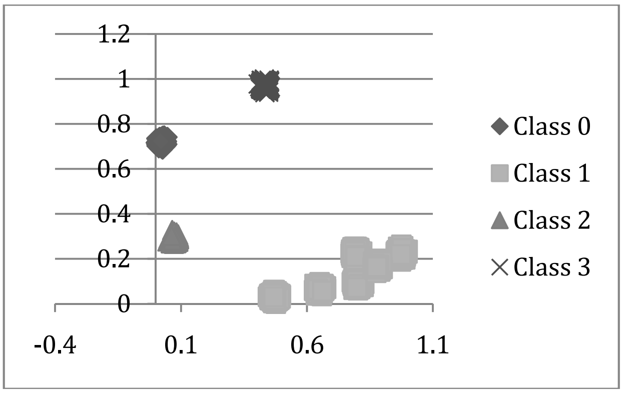

Fig. 5. Data set in Fig. 4 divided into k=4 clusters

C. *Explanatory Clustering*

The final wrinkle to explain why the validity measures give such widely different results is that the validity measures, like clustering itself, are merely ways of looking at the data and interpreting it. While all validity measures agree that separation and compactness are desirable and more separation and compactness will predict a better clustering, each method defines separation and compactness in its own way.

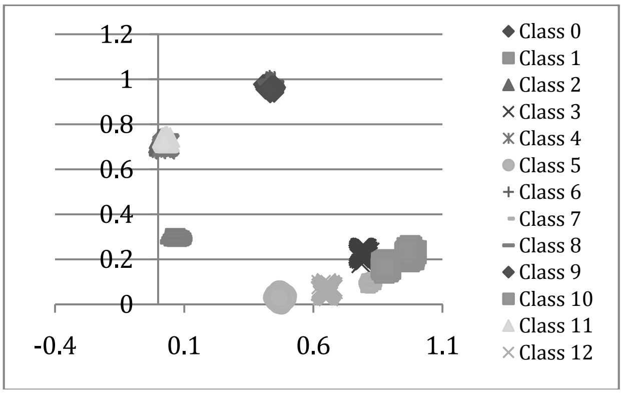

Fig. 6. Data set in Fig. 4 divided into k=13 clusters

Again, to use the ten class data set above as an example, two of the validation algorithms liked the four-cluster partitioning better than the 13 cluster partition. To the human eye, such a partition makes sense—and once it has been pointed out, it may even make more sense than the obvious nine-cluster partition which is closest to the "actual" ten classes. In such ambiguous situations, different priorities assigned to separation and compactness, and different methods for their computation will almost inevitably yield different recommendations.

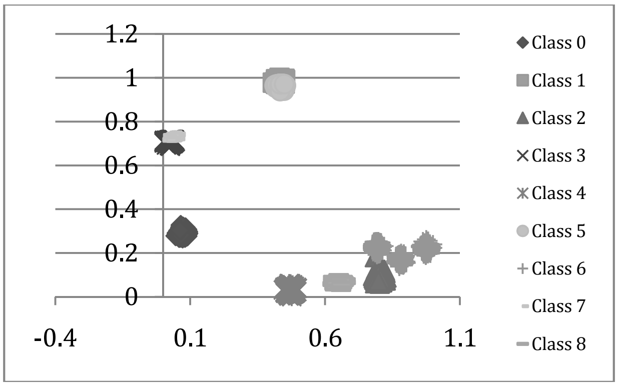

Fig. 7. Data set in Fig. 4 divided into k=9 clusters

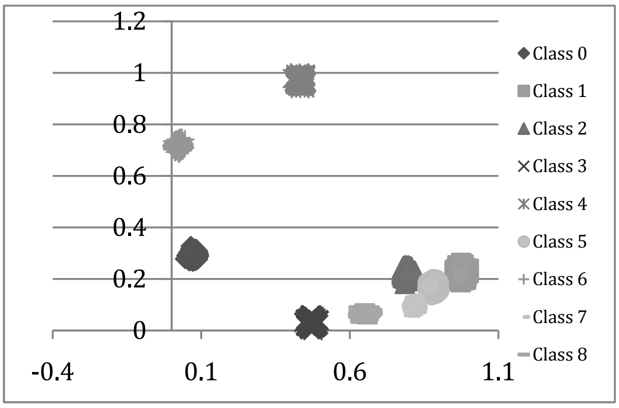

Fig. 8. Data set in Fig. 4 divided into k=9 clusters (random initialization with 100 runs)

V. CONCLUSIONS

In terms of choosing an effective validation technique for predicting the appropriate value of k, it seems that the SS method works better than many of its competitors. Silhouette is a good alternative, and, surprisingly, CH and DB both make only moderate showings compared to the other two. Dunn, PB, and PBM performed quite poorly.

However, the results of these tests need to be qualified by the nature of the validity measure. Validity measures provide a better analysis of data than the mere compactness criteria included in the k-means test, by also taking separation into account and trying to normalize the results so that different clusterings can be compared. However, they are limited by the clusterings that are given. If the k-means does not find the optimal solution for any particular given k, the validity measures will be providing sub-optimal results. And the different means that measures have for evaluating compactness and separation will also inevitably result in varying results.

More work needs to be done on the use of validity measures as interpretive tools for understanding data that is either too difficult to visualize or apparently ambiguous. Investigations into cluster validity indexes should move into more complex issues than “right k” or “wrong k.” How the validity measure determined a particular k, and what that might mean about the different ways in which the data might be understood or structured, and how those predictions relate to the predictions and measures provided by other validity measures are questions that need to be asked.

REFERENCES

[1] Jain, A. K. 2010. Data Clustering: 50 Years Beyond K-means. *Pattern Recognition Letters* 31(8): 651–666.

[2] Jain, A. K.; Murty, M. N.; and Flynn, P. J. 1999. Data Clustering: A Review. *ACM Computing Surveys* 31(3): 264–323.

[3] Halkidi, M.; Batistakis, Y.; and Vazirgiannis, M. 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems* 17(2/3): 107–145.

[4] Milligan, G. W. and Cooper, M. C. 1985. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* 50(2): 159–179.

[5] Vendramin, L.; Campello, R. J. G. B.; and Hruschka, E. R. 2010. Relative Clustering Validity Criteria: A Comparative Overview. *Statistical Analysis and Data Mining* 3(4): 209–235.

[6] Zhao, Q.; Xu, M.; and Franti, P. 2009. Sum-of-Squares Based Cluster Validity Index and Significance Analysis. *Proceedings of the International Conference on Adaptive and Natural Computing Algorithms (ICANNGA 2009), Lecture Notes in Computer Science* 5495: 313–322.

[7] Dunn, J.C. 1974. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics* 4(1): 95–104.

[8] Davies, D. and Bouldin, D. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(2): 224–227.

[9] Rousseeuw, P. J. 1987. Silhouettes: A Graphic Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* 20(1): 53–65.

[10] Calinski, R.B. and Harabasz, J. 1974. A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods* 3(1): 1–27.

[11] Milligan, Glenn W. 1981. A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis. *Psychometrika* 46(2): 187–199.

[12] Pakhira, M. K.; Bandyopadhyay, S.; and Maulik, U. 2004. Validity Index for Crisp and Fuzzy Clusters. *Pattern Recognition* 37(3): 487–501.

[13] Maitra, R. and Melnykov, V. 2010. Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms. *Journal of Computational and Graphical Statistics* 19(2): 354–376.

[14] Lloyd, S. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory* 28(2): 129–136.

[15] Gonzalez, T. 1985. Clustering to Minimize the Maximum Intercluster Distance. *Theoretical Computer Science* 38(2/3): 293–306.