K-means Clustering versus Validation Measures: A Data Distribution Perspective

Hui Xiong
Rutgers University
hui@rbs.rutgers.edu

Junjie Wu
Tsinghua University
wujj@em.tsinghua.edu.cn

Jian Chen
Tsinghua University
jchen@mail.tsinghua.edu.cn

ABSTRACT

K-means is a widely used partitional clustering method. While there are considerable research efforts to characterize the key features of K-means clustering, further investigation is needed to reveal whether and how the data distributions can have the impact on the performance of K-means clustering. Indeed, in this paper, we revisit the K-means clustering problem by answering three questions. First, how the "true" cluster sizes can make impact on the performance of K-means clustering? Second, is the entropy an algorithm-independent validation measure for K-means clustering? Finally, what is the distribution of the clustering results by K-means? To that end, we first illustrate that K-means tends to generate the clusters with the relatively uniform distribution on the cluster sizes. In addition, we show that the entropy measure, an external clustering validation measure, has the favorite on the clustering algorithms which tend to reduce high variation on the cluster sizes. Finally, our experimental results indicate that K-means tends to produce the clusters in which the variation of the cluster sizes, as measured by the Coefficient of Variation (CV), is in a specific range, approximately from 0.3 to 1.0.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.5.3 [**Pattern Recognition**]: Clustering

General Terms

Algorithms, Experimentation

Keywords

K-means Clustering, Coefficient of Variation (CV), Entropy

1. INTRODUCTION

Cluster analysis [9] provides insight into the data by dividing the objects into groups (clusters) of objects, such that objects in a cluster are more similar to each other than to objects in other clusters. K-means [15] is a well-known and widely used partitional clustering method. In the literature, there are considerable research efforts to characterize the key features of K-means clustering. Indeed, people have identified some characteristics of data that may strongly affect the K-means clustering analysis including high dimensionality, the size of the data, the sparseness of the data, noise and outliers in the data, types of attributes and data sets, and scales of attributes [20]. However, further investigation is needed to reveal whether and how the data distributions can have the impact on the performance of K-means clustering. Along this line, in this paper, we revisit K-means clustering by answering three questions.

1. How can the distribution of "true" cluster sizes make impact on the performance of K-means clustering?
2. Is the entropy an algorithm-independent validation measure for K-means clustering?
3. What is the distribution of the clustering results by K-means?

The answers to these questions can guide us for the better understanding and the use of K-means. This is noteworthy since, for document data, K-means has been shown to perform as well or better than a variety of other clustering techniques and has the appealing computational efficiency [19, 22]. To this end, we first illustrate that K-means clustering tends to generate the clusters with the relatively uniform distribution on the cluster sizes. Also, we show that the entropy measure, an external clustering validation measure, has the favorite on the clustering algorithms, such as K-means, which tend to reduce the variation on the cluster sizes. In other words, entropy is not an algorithm-independent validation measure.

In addition, we have conducted extensive experiments on a number of real-world data sets from various different application domains. Our experimental results show that K-means tends to produce the clusters in which the variation of the cluster sizes is in a specific range. This data variation is measured by the Coefficient of Variation (CV) [2]. The CV, described in more detail later, is a measure of dispersion of a data distribution and is a dimensionless number that allows comparison of the variation of populations that have significantly different mean values. In general, the larger the CV value is, the greater the variability in the data.

Indeed, as shown in our experimental results, for the data sets with high variation on the "true" cluster size (e.g. CV $>$ 1.0), K-means reduces this variation in the resulting cluster sizes to less than 1.0. Meanwhile, for the data sets with low variation on the "true" cluster size (e.g. CV $<$ 0.3), K-means

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

increases the variation slightly in the resulting cluster sizes to greater than 0.3 . In other words, for these two cases, K-means produces the clustering results which are away from the "true" cluster distributions.

2. RELATED WORK

People have investigated K-means clustering from various perspectives. Many data factors, which may strongly affect the performance of K-means, have been identified and addressed. In the following, we highlight some results which are mostly related to the main theme of this paper.

First, people have studied the impact of the high dimensionality on the performance of K-means and found that the traditional Euclidean notion of proximity is not very effective for K-means on high-dimensional data sets, such as gene expression data sets and document data sets. To meet this challenge, one research direction is to employ dimensionality reduction techniques, such as Multidimensional Scaling (MDS) or Singular Value Decomposition(SVD). Another direction is to redefine the notions of proximity, e.g., by the Shared Nearest Neighbors (SNN) similarity [10].

Second, many clustering algorithms that work well for small or medium-size data sets are unable to handle larger data sets. Along this line, a discussion of scaling K-means clustering to large data sets is provided by Bradley et al. [1]. Also, Ghosh [5] discussed the scalability of clustering methods in depth and a more broad discussion of specific clustering techniques can be found in [16].

Third, outliers and noise in the data can also degrade the performance of clustering algorithms, especially for prototype-based algorithms such as K-means. There has been serval techniques designed for handling this problem. For example, DBSCAN automatically classifies low-density points as noise and removes them from the clustering process [4]. Chameleon [12], SNN density-based clustering [3], and CURE [6] explicitly deal with noise and outliers during the clustering process.

Finally, the researchers have identified some other data factors, such as the types of attributes, the types of data sets, and scales of attributes, which may have the impact on the performance of K-means clustering. However, in this paper, we target on understanding the impact of the distribution of the "true" cluster size on the performance of K-means clustering and the cluster distribution of the clustering results by K-means. Also, we investigate the relationship between K-means and the entropy measure.

3. THE JOINT EFFECT OF K-MEANS CLUSTERING AND THE ENTROPY MEASURE

In this section, we illustrate the effect of K-means clustering on the distribution of the cluster sizes, and show the relationship between the entropy measure and K-means.

K-means [15] is a prototype-based, simple partitional clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids (a cluster centroid is typically the mean of the points in the cluster). The clustering process of K-means is as follows. First, k initial centroids are selected, where k is specified by the user and indicates the desired number of clusters. Every point in the data is then assigned to the closest centroid, and each collection of points assigned to a centroid forms a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. This process is repeated until no point changes clusters.

In general, there are two kinds of clustering validation techniques, which are based on external criteria and internal criteria respectively. Entropy is a commonly used external validation measures for K-means clustering [19, 22]. As an external criteria, entropy uses external information — class labels in this case. Indeed, entropy measures the purity of the clusters with respect to the given class labels. Thus, if every cluster consists of objects with only a single class label, the entropy is 0. However, as the class labels of objects in a cluster become more varied, the entropy value increases.

To compute the entropy of a set of clusters, we first calculate the class distribution of the objects in each cluster, i.e., for each cluster j we compute p_{ij}, the probability that a member of cluster j belongs to class i. Given this class distribution, the entropy of cluster j is calculated using the standard entropy, $E_j = -\sum_i p_{ij} log(p_{ij})$, where the sum is taken over all classes and the log is log base 2. The total entropy, $E = \sum_{j=1}^{m} \frac{n_j}{n} E_j$, for a set of clusters is computed as the weighted sum of the entropies of each cluster, where n_j is the size of cluster j, m is the number of clusters, and n is the number of all data points.

In a similar fashion, we can compute the purity of a set of clusters. First, we calculate the purity of each cluster. For each cluster j, we have the purity $P_j = \max_i(n_j^i)/n_j$, where n_j^i is the number of objects in cluster j with class label i. In other words, P_j is the fraction of the overall cluster size that the largest class of objects assigned to that cluster represents. The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and is given as $Purity = \sum_{j=1}^{m} \frac{n_j}{n} P_j$, where n_j is the size of cluster j, m is the number of clusters, and n is the number of all data points. In general, we believe that the larger the value of purity, the better the clustering solution is.

3.1 Dispersion Degree of Data Distributions

Before we describe the joint effect of K-means clustering and the entropy measure, we first introduce Coefficient of Variation (CV) [2], which is a measure of dispersion for a data distribution. CV is defined as the ratio of the standard deviation to the mean. Given a set of data objects $X = \{x_1, x_2, \ldots, x_n\}$, we have CV $= \frac{s}{\bar{x}}$, where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$.

Please note that there are some other statistics, such as standard deviation and skewness, which can also be used to characterize the dispersion degree of data distributions. However, the standard deviation has no scalability; that is, the dispersion degree of the original data and the stratified sample data is not equal as indicated by standard deviation, which does not agree with our intuition. Meanwhile, skewness cannot catch the dispersion in the situation that the data are symmetric but do have high variance. Indeed, CV is a dimensionless number that allows comparison of the variation of populations that have significantly different mean values. In general, the larger the CV value, the greater the variability is in the data.

3.2 The Effect of K-means Clustering on the Distribution of the Cluster Sizes

In this section, we illustrate the effect of K-means clustering on the distribution of the cluster sizes.

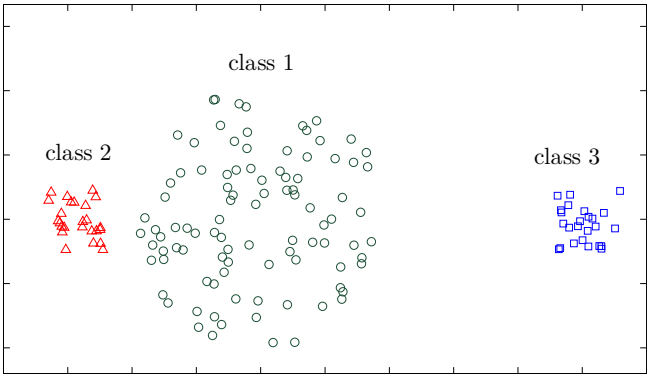

Figure 1: Clusters before K-means Clustering.

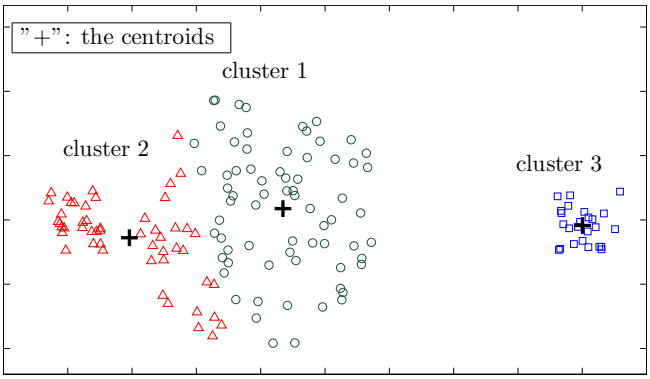

Figure 2: Clusters after K-means Clustering.

Figure 1 shows a sample data set with three "true" clusters. The numbers of points in Cluster 1, 2 and 3 are 96, 25 and 25, respectively. In this data, Cluster 2 is much closer to Cluster 1 than Cluster 3. Figure 2 shows the clustering results by K-means on this data set. As can be seen, three natural clusters could not be identified exactly. One observation is that Cluster 1 is broken: part of Cluster 1 is merged with Cluster 2 as new Cluster 2 and the rest of cluster 1 forms new Cluster 1. However, the size distribution of the resulting two clusters is more uniform now. This is called the "**uniform effect**" of K-means on "true" clusters with different sizes. Another observation is that Cluster 3 is precisely identified by K-means. It is due to the fact that the objects in Cluster 3 are far away from Cluster 1. In other words, the uniform effect has been dominated by the large distance between two clusters. From the above, we can notice that the uniform effect of K-means clustering on "true" clusters with different sizes does exist. We will further illustrate this in our experimental section.

Table 1: A Sample Document Data Set.

A Sample Document Data Set
Sports, Sports
Entertainment, Entertainment
Foreign, Foreign, Foreign, Foreign, Foreign
Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro
Politics
CV=1.1187

3.3 The Limitations of the Entropy Measure

In our practice, we have observed that the entropy measure tends to favor clustering algorithms, such as K-means, which produce clusters with relatively uniform sizes. We call this the "**biased effect**" of the entropy measure. To illustrate this, we created the sample data set as shown in Table 1. This data set consists of 42 documents with 5 class labels. In other words, there are five "true" clusters in this sample data. The CV value of the cluster sizes of these five "true" clusters is 1.1187 as presented in the table.

Table 2: Two Clustering Results.

Document Clustering		
Clustering I	1: Sports Sports Sports Sports Sports Sports Sports Sports	
	2: Sports Sports Sports Sports Sports Sports Sports Sports	CV=0.4213
	3: Sports Sports Sports Sports Sports Sports Sports Sports	Purity=0.929
	4: Metro Metro Metro Metro Metro Metro Metro Metro Metro Metro	Entropy=0.247
	5: Entertainment Entertainment Foreign Foreign Foreign Foreign Politics	
Clustering II	1: Sports Foreign	
	2: Entertainment Entertainment	CV=1.2011
	3: Foreign Foreign Foreign	Purity=0.952
	4: Metro Metro Metro Metro Metro Metro Metro Metro Metro Metro Foreign	Entropy=0.259
	5: Politics	

For this sample document data set, we assume that we have two clustering results by different clustering algorithms as shown in Table 2. In the table, we can observe that the first clustering result has five clusters with relatively uniform sizes. This is also indicated by the CV value, which is 0.4213. In contrast, for the second clustering result, the CV value of the cluster sizes is 1.2011. This indicates that the five clusters have widely different cluster sizes for the second clustering scheme. Certainly, according to entropy, clustering result I is better than clustering result II (This result is due to the fact that the entropy measure more heavily penalizes a large impure cluster.) However, if we look at five "true" clusters carefully, we find that the second clustering results are much closer to the "true" cluster distribution and the first clustering results are actually away from the "true" cluster distribution. This is also reflected by the CV values. The CV value of five cluster sizes in the second clustering results is closer to the CV value of five "true" cluster sizes.

Finally, in Table 2, we can also observe that the purity of the second clustering results is better than that of the first clustering results. Indeed, this is contradict to the results by the entropy measure. In summary, this example illustrates that the entropy measure has the favorite on the algorithms, such as K-means, which produce clusters with relatively uniform sizes. In other words, if the entropy measure is used for validating the K-means clustering, the validation results can be misleading.

4. EXPERIMENTAL EVALUATION

In this section, we present experimental results to show the impact of data distributions on the performance of K-

Table 4: Some Characteristics of Experimental Data Sets.

Data Set	Source	# of Objects	# of Features	# of Classes	Min Class Size	Max Class Size	CV_0
Document Data Set							
fbis	TREC	2463	2000	17	38	506	0.961
hitech	TREC	2301	126373	6	116	603	0.495
sports	TREC	8580	126373	7	122	3412	1.022
tr23	TREC	204	5832	6	6	91	0.935
tr45	TREC	690	8261	10	14	160	0.669
la2	TREC	3075	31472	6	248	905	0.516
ohscal	OHSUMED-233445	11162	11465	10	709	1621	0.266
re0	Reuters-21578	1504	2886	13	11	608	1.502
re1	Reuters-21578	1657	3758	25	10	371	1.385
k1a	WebACE	2340	21839	20	9	494	1.004
k1b	WebACE	2340	21839	6	60	1389	1.316
wap	WebACE	1560	8460	20	5	341	1.040
Biomedical Data Set							
LungCancer	KRBDSR	203	12600	5	6	139	1.363
Leukemia	KRBDSR	325	12558	7	15	79	0.584
UCI Data Set							
ecoli	UCI	336	7	8	2	143	1.160
page-blocks	UCI	5473	10	5	28	4913	1.953
pendigits	UCI	10992	16	10	1055	1144	0.042
letter	UCI	20000	16	26	734	813	0.030

Table 5: Experimental Results for Real-world Data Sets.

		Standard Deviation of Sizes		Coefficient of Variation of Sizes			
Data Set	Average of Sizes	STD_0	STD_1	CV_0	CV_1	DCV=CV_0-CV_1	Entropy
fbis	145	139	80	0.96	0.55	0.41	0.345
hitech	384	190	140	0.50	0.37	0.13	0.630
sports	1226	1253	516	1.02	0.42	0.60	0.190
tr23	34	32	14	0.93	0.42	0.51	0.418
tr45	69	46	30	0.67	0.44	0.23	0.329
la2	513	264	193	0.52	0.38	0.14	0.401
ohscal	1116	297	489	0.27	0.44	-0.17	0.558
re0	116	174	45	1.50	0.39	1.11	0.374
re1	66	92	22	1.39	0.33	1.06	0.302
k1a	117	117	57	1.00	0.49	0.51	0.342
k1b	390	513	254	1.32	0.65	0.66	0.153
wap	78	81	39	1.04	0.49	0.55	0.313
LungCancer	41	55	26	1.36	0.63	0.73	0.332
Leukemia	46	27	17	0.58	0.37	0.21	0.511
ecoli	42	49	21	1.16	0.50	0.66	0.326
page-blocks	1095	2138	1029	1.95	0.94	1.01	0.146
pendigits	1099	46	628	0.04	0.57	-0.53	0.394
letter	769	23	440	0.03	0.57	-0.54	0.683
Min	34	23	14	0.03	0.33	-0.54	0.146
Max	1226	2138	1029	1.95	0.94	1.11	0.683
Parameters used in CLUTO: -clmethod=rb -sim=cos -crfun=i2 -niter=30							

Table 3: Some Notations.

CV_0: the CV value on the cluster sizes of the "true" clusters
CV_1: the CV value on the cluster sizes of the clustering results
DCV: the change of CV values before and after K-means clustering

means clustering. Specifically, we demonstrate: (1) the effect of the "true" cluster sizes on K-means clustering; and (2) the effect of the entropy measure on the K-means clustering results.

4.1 The Experimental Setup

Experimental Tool. In our experiments, we used the implementation of K-means in CLUTO [11]. For all the experiments, the cosine similarity is used in the objective function for K-means. Finally, please note that some notations used in our experiments are shown in Table 3.

Experimental Data Sets. For our experiments, we used a number of real-world data sets that were obtained from different application domains. Some characteristics of these data sets are shown in Table 4. In the table, "# of Classes" indicates the number of "true" clusters.

Document Data Sets. The `fbis` data set was from the Foreign Broadcast Information Service data of the TREC-5 collection [21]. The `hitech` and `sports` data sets were derived from the San Jose Mercury newspaper articles that were distributed as part of the TREC collection (TIPSTER Vol. 3). Data sets `tr23` and `tr45` were derived from the TREC-5[21], TREC-6 [21], and TREC-7 [21] collections. The `la2` data set was part of the TREC-5 collection [21] and contains news articles from the Los Angeles Times. The `ohscal` data set was obtained from the OHSUMED collection [8], which contains documents from various biological sub-fields. The data sets `re0` and `re1` were from Reuters-21578 text categorization test collection Distribution 1.0 [13]. The data sets `k1a` and `k1b` contain exactly the same set of documents but they differ in how the documents were assigned to different classes. In particular, `k1a` contains a finer-grain categorization than that contained in `k1b`. The data set `wap` was from the WebACE project (WAP) [7]; each document corresponds to a web page listed in the subject hierarchy of Yahoo!. For all document clustering data sets, we used a stop-list to remove common words, and the words were stemmed using Porter's suffix-stripping algorithm [18].

Biological Data Sets. `LungCancer` and `Leukemia` data sets were from the Kent Ridge Biomedical Data Set Repository (KRBDSR) which is an online repository of high dimensional features [14]. The `LungCancer` data set consists of samples of lung adenocarcinomas, squamous cell lung carcinomas, pulmonary carcinoid, small-cell lung carcinomas and normal lung described by 12600 genes. The `Leukemia` data set contains 6 subtypes of pediatric acute lymphoblastic leukemia samples and 1 group samples that do not fit in any of the above 6 subtypes, and each is described by 12558 genes.

UCI Data Sets [17]. The `ecoli` data set is about the information of cellular localization sites of proteins. The `page-blocks` data set contains the information of 5-type blocks of the page layout of a document that has been detected by a segmentation process. The `pendigits` and `letter` data sets contain the information of handwritings. The former is the numeric information of 0-9, and the latter letter information of A-Z.

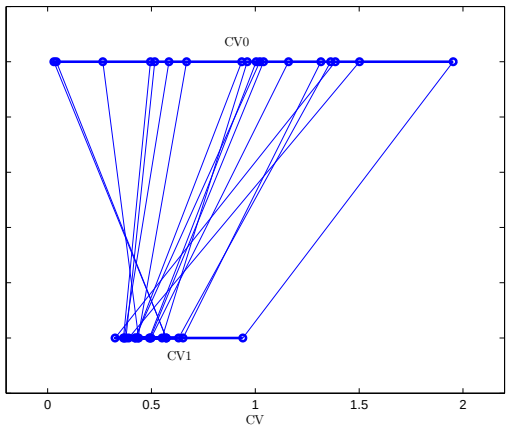

Figure 3: The Distributions of CV Values before and after K-means Clustering.

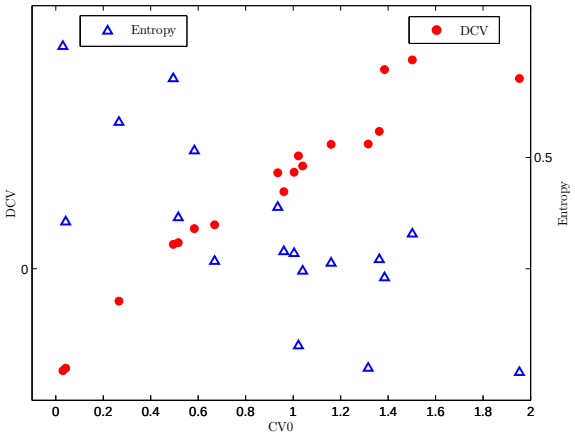

Figure 4: Illustration of the "Biased Effect" of Entropy on All the Experimental Data Sets.

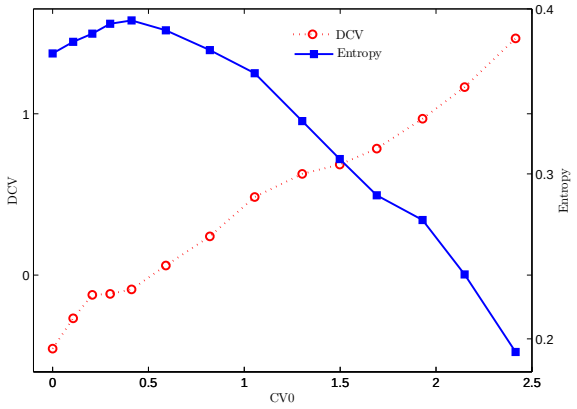

Figure 5: Illustration of the "Biased Effect" of Entropy Using Sample Data Sets from "Pendigits".

4.2 The Effect of the "True" Cluster Sizes on K-means Clustering

Here, we illustrate the effect of the "true" cluster sizes on the results of K-means clustering. In our experiment, we first used K-means to cluster the input data sets, and then computed the CV values for the "true" cluster distribution of the original data and the cluster distribution of the clustering results. The number of clusters k was set as the "true" cluster number for the purpose of comparison.

Table 5 shows the experimental results on various real-world data sets. As can be seen, for the data sets with large CV_0, K-means tends to reduce the variation on the cluster sizes of the clustering results as indicated by CV_1. This result indicates that, for data sets with high variation on the cluster sizes of "true" clusters, the uniform effect of K-means is dominant; that is, K-means tends to reduce the variation on the cluster sizes in this case.

Another observation is that, for data sets with low CV_0 values, K-means increases the variation on the cluster sizes of the clustering results slightly as indicated by the corresponding CV_1 values. This result indicates that, for data sets with very low variation on the cluster sizes of "true" clusters, the uniform effect of K-means is not significant. Other factors, e.g., the variant shapes, densities, or the centroid distances between the "true" clusters, tend to be the dominant factors instead.

Indeed, Figure 3 shows the link relationships between CV_0 and CV_1 for all the experimental data sets listed in Table 4, and there is a link between CV_0 and CV_1 for every data set. A very interesting observation is that, while the range of CV_0 is between 0.03 and 1.95, the range of CV_1 is restricted into a much smaller range from 0.33 to 0.94. Thus we empirically have the interval of CV_1 values: [0.3, 1.0].

4.3 The Effect of the Entropy Measure on the K-means Clustering Results

In this subsection, we present the effect of the entropy measure on the K-means clustering results. Figure 4 shows the plot of entropy values for all the experimental data sets in Table 4. A general trend can be observed is that while the differences in CV values before and after clustering increase as the increase of CV_0 values, the entropy values tend to

decrease. In other words, there is a disagreement between DCV and the entropy measure on evaluating the clustering quality. The entropy measure indicates better quality, but DCV shows that the distributions of clustering results are away from the distributions of "true" clusters. This indicates worse clustering quality. The above observation agrees with our analysis in Section 3 that the entropy measure has a biased effect on K-means.

To strengthen the above observations, we also generated two groups of synthetic data sets from two real-world data sets: `pendigits` and `letter`. To generate data sets from `pendigits`, we applied the following sampling strategy: 1) We first sampled the original data set to get a sample with 10 "true" clusters, each of which contains $1000, 100, 100, \cdots, 100$ objects, respectively. Then 2) we did random sampling on the biggest cluster and merged the samples with all the other objects in the rest 9 clusters to form a data set. We gradually reduced the sample size to 100, thus obtained various data sets with decreasing dispersion degrees. On the other hand, in order to have data sets with increasing dispersion degrees, 3) we did random, stratified sampling to the 9 smaller clusters, and merged the samples with the rest 1000 objects to form a data set. We gradually reduced the sample size for each of the 9 clusters to 30, thus got a series of data sets with increasing dispersion degrees. A similar sampling strategy was also applied to `letter`. Note that for each dispersion degree we did sampling 10 times and output the average values as the sampling results.

Figure 5 shows the corresponding plot of the entropy values for the synthetic data sets derived from the `pendigits` data set. A similar trend has been observed; that is, the entropy values and the DCV values do not agree with each other for clustering validation as the increase of CV_0 values. Due to the page limit, we have omitted a similar plot for the second group of synthetic data sets derived from the `letter` data set.

5. CONCLUSIONS

In this paper, we illustrate the relationship between K-means and the "true" cluster sizes as well as the entropy measure. Our experimental results demonstrate that K-means tends to reduce the variation on the cluster sizes if the variation of the "true" cluster sizes is high and increase the variation on the cluster sizes if the variation of the "true" cluster sizes is very low. In addition, we found that, no matter what are the CV values of the "true" cluster sizes, the CV values of the clustering results are typically located in a much smaller range from 0.3 to 1.0. Finally, we observed that many "true" clusters were disappeared in the clustering results if K-means is applied for data sets with high variation on the "true" cluster sizes; that is, K-means produces the clustering results which are far away from the "true" cluster distribution. This is actually contradicted by the entropy measure, since the entropy values are usually very low for the data sets with high variation on the "true" cluster sizes. In other words, the entropy measure is not an algorithm-independent clustering validation measure and has the favorite on K-means.

6. REFERENCES

[1] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *KDD*, pages 9–15, 1998.

[2] M. DeGroot and M. Schervish. *Probability and Statistics*. Addison Wesley; 3rd edition, 2001.

[3] L. Ertoz, M. Steinbach, and V. Kumar. A new shared nearest neighbor clustering algorithm and its applications. In *SDM Workshop on Clustering High Dimensional Data and its Applications*, 2001.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[5] J. Ghosh. *Scalable Clustering Methods for Data Mining, Handbook of Data Mining*. Lawrence Ealbaum Assoc, 2003.

[6] S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. In *ACM SIGMOD*, pages 73–84, 1998.

[7] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. Webace: A web agent for document categorization and exploration. In *Proc. of the 2nd Intl. Conf. on Autonomous Agents*, 1998.

[8] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR*, pages 192–201, 1994.

[9] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1998.

[10] R. Jarvis and E. Patrick. Clusering using a similarity measure based on shared nearest neighbors. *IEEE Trans. on Computers*, C-22(11):1025–1034, 1973.

[11] G. Karypis. In *http://www-users.cs.umn.edu/ karypis/cluto/*.

[12] G. Karypis, E.-H. Han, and V. Kumar. Chameleon: A hierarchical clustering algorithm using dynamic modeling. *IEEE Computer*, 32(8):68–75, August 1999.

[13] D. Lewis. Reuters-21578 text categorization text collection 1.0. In *http://www.research.att.com/ lewis*.

[14] J. Li and H. Liu. In *http://sdmc.i2r.a-star.edu.sg/rp/*.

[15] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume I, Statistics*. University of California Press, September 1967.

[16] F. Murtagh. *Clustering Massive Data Sets, Handbook of Massive Data Sets*. Kluwer, 2000.

[17] D. Newman, S. Hettich, C. Blake, and C. Merz. Uci repository of machine learning databases, 1998.

[18] M. F. Porter. An algorithm for suffix stripping. In *Program, 14(3)*, 1980.

[19] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, August 2000.

[20] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[21] TREC. In *http://trec.nist.gov*.

[22] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. *Machine Learning*, 55(3):Pages: 311–331, June 2004.