EFFICIENT UNIMODALITY TEST IN CLUSTERING BY SIGNATURE TESTING

Mahdi Shahbaba and Soosan Beheshti

Ryerson University
Department of Computer Science and Electrical Engineering
350 Victoria Street Toronto, ON M5B 2K3
mshahbab@ryerson.ca, soosan@ee.ryerson.ca

ABSTRACT

This paper provides a new unimodality test with application in hierarchical clustering methods. The proposed method denoted by signature test (Sigtest), transforms the data based on its statistics. The transformed data has much smaller variation compared to the original data and can be evaluated in a simple proposed unimodality test. Compared with the existing unimodality tests, Sigtest is more accurate in detecting the overlapped clusters and has a much less computational complexity. Simulation results demonstrate the efficiency of this statistic test for both real and synthetic data sets.

Index Terms— Clustering, Statistic test, Unimodality test

1. INTRODUCTION

Data clustering is an unsupervised learning algorithm for grouping similar data samples [1]. The family of clustering methods only rely on data itself when *a priori* knowledge about the labels and classes is not available. One main challenge in this type of clustering is finding the correct number of clusters involved [2]. Hierarchical clustering algorithms answer this problem by using a cluster splitting criteria. These methods test a null hypothesis for distribution of a single cluster and split the dataset until all estimated clusters pass the test. An improper statistical test for splitting criterion will lead to an incorrect estimation of the number of clusters. This problem generally is caused due to the lack of a universal statistic test for all types of clusters. Statistical tests in these approaches are in form of unimodality test. Examples of these unimodality tests are Anderson-Darling[3], Kolmogorov-Smirnov[4] and dip test[5].

In this paper, we provide a new splitting criterion for unimodality that can also be used in hierarchical clustering algorithms. Our proposed criterion relies on compressing the data based on its statistics and leads to minimizing the data variation by transforming the data. The transformation is denoted in form of signatures. The data signatures for statistical data plays analogous role to the data sparse transformation for sparse signals, i.e., it transforms the data such that the signatures of statistics will be extracted from the data itself. One of the advantages of this statistical test, denoted by Sigtest, is its robustness for recognizing the highly overlapped clusters compared with the state of the art unimodality tests.

2. RELATED WORK

The correct number of clusters is a crucial parameter which either is available before clustering or should be estimated by clustering methods. X-means algorithm is one of the first hierarchical clustering methods which relies on Bayesian Information criterion (BIC) for cluster splitting [6], [7]. This method only recognizes spherical Gaussian clusters and splits clusters with non-spherical distribution.

G-means benefits from Anderson-Darling statistic test (AD) for examining the Gaussianity of clusters and similar to X-means it is a wrapper around K-means algorithm. In contrast to X-means, G-means can deal with any distribution from Gaussian family [8], [9]. Employing the Expectation Maximization algorithm (EM), PG-means clustering can deal with overlapped clusters better than G-means [10]. PG-means projects model and all of the dataset on several random directions, and then using Kolmogorov-Smirnov (KS) decides whether model and dataset are matched for each projection.

Dip-means clustering is constructed based on the Hartigan's dip test of unimodality [11]. According to this clustering method, each sample is a viewer with different distance values from other samples. Using dip test, distribution of the distance values should be examined for unimodality. If all viewers pass the unimodality test then null hypothesis of having a single cluster will be approved. Otherwise, a model with more than one cluster should be considered for the samples. This method is a wrapper around K-means, which can also work with kernel K-means to detect arbitrary shape clusters.

Both dip-means and G-means rely on statistical tests for cluster splitting, but accuracy of these criteria remains a concern for the case of overlapped clusters. Our proposed statistic test defines probabilistic bounds on signature of a single cluster and employs it as a reference for comparison with any given cluster. The accuracy of proposed Sigtest will be compared with dip-test, KS and AD using synthetic dataset. Also

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

modified versions of dip-means and G-means based on Sigtest are evaluated using real benchmark datasets from UCI repository database.

3. PROBLEM STATEMENT

Let $y = [y_1, y_2, \cdots, y_N]^T$ be a vector of n observations; where $y_i \in R$ is generated from an unknown distribution. There, exist the following possible hypotheses for unimodality test on y:
•H_0: y is sampled from a unimodal distribution.
•H_1: y is not sampled from a unimodal distribution.
where, acceptance of each hypothesis will affect the cluster splitting criterion of the model.

3.1. Application of Unimodality Test in Clustering

Hierarchical clustering methods rely on cluster splitting criteria to recognize single clusters in a given dataset. In Fig.1, C_i represents the data which should be checked for splitting at the i^{th} stage of clustering. If the criterion accepts splitting ($H_1 \equiv split = 1$), C_i will be split into two new clusters C_{i1} and C_{i2}, otherwise ($H_0 \equiv split = 0$) it remains as one cluster. In the following stages, the checking procedure continues for all new clusters (if any) until H_0 is valid for all clusters.
In this clustering, the available data (x) has dimension of d, however, splitting criterion is usually based on a transformation of d-dimensional data to a one-dimensional data. The first step in using the criterion is transforming x to y:

$$y = f(x) \tag{1}$$

where f is the transformation of x to one-dimensional y.

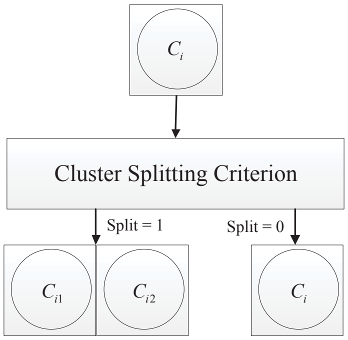

Fig. 1: Hierarchical clustering and data splitting.

4. UNIMODALITY SIGNATURE TEST (SIGTEST)

In this section, we define a new unimodality test for hierarchical clustering which relies on data signature and probabilistic bounds on its distribution. Signature of a data is a function of data that compresses the data in a proper transformation[12]. To illustrate an example of data signature, Fig.2 shows 1000 randomly generated Gaussian samples from $\mathcal{N}(0,1)$ for 100 runs. As the figure illustrates, while samples themselves (in the top plot) vary between $\pm 3\sigma$, the middle plot which is sorted version of the same samples is a transformation of the top plot in a much more compact form, i.e., the variance of the sorted version is smaller than $\frac{1}{10}$ of the original variance. Therefore, we can define a probabilistic confidence region around the dense area, with a very small variance, and use it as a signature for unimodality.

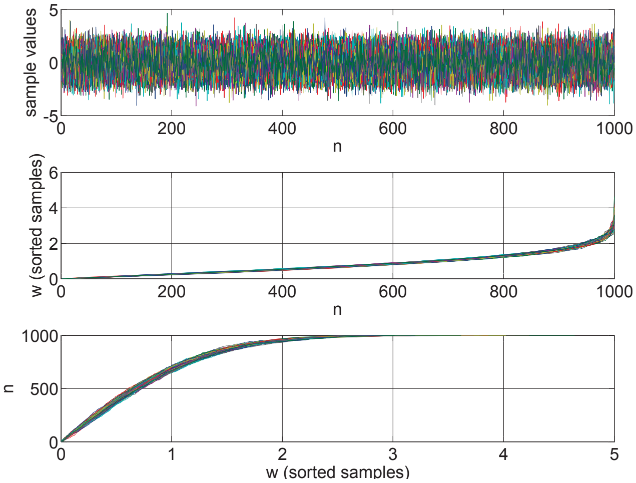

Fig. 2: Top figure: 100 runs of a zero mean and unit variance Gaussian distribution with length 1000. Middle: sorted absolute values of the top figure. Bottom: The same middle figure with swapped axes.

4.1. Probabilistic Bounds on Signature of Unimodal Distribution

We propose two signatures $g_1(w_n)$ and $g_2(w_n)$ for unimodality test. Let $w = [w_1, w_2, \cdots, w_N]$ represent a random variable that is sorted absolute value of samples from a unimodal distribution. The first suggested signature as it was shown in Fig. 2 can be the sorted version of data itself:

$$g_1(w_n) = w_n \tag{2}$$

This signature has the following expected value and variance:

$$E[w_n] = F(w_n) \tag{3}$$

$$var[w_n] = \frac{1}{N} F(w_n)(1 - F(w_n)) \tag{4}$$

where $F(w_n)$ is the cumulative distribution function (cdf) of w_n. Details of calculation of this expected value and variance are provided in [12]. Another proposed signature $g_2(w_n)$ is:

$$g_2(w_n) = \frac{1}{n} \sum_{j=1}^{n} w_j \tag{5}$$

Using (3), the expected value of $g_2(w_n)$ is:

$$E[g_2(w_n)] = E[w_n] \tag{6}$$

and using (4), it can be shown that the variance of $g_2(w_n)$ is bounded as follows[1]:

$$var[g_2(w_n)] \leq var[w_n] \tag{9}$$

Note that both of the above signatures have very small variances compared to their original distribution similar to what is shown in Fig. 2. The tight boundaries of each signature as a function of their indexes are denoted by $U(n)$ and $L(n)$:

$$\begin{aligned} U(n) &= E[g_i(w_n)] + \gamma\sqrt{var[g_i(w_n)]} \\ L(n) &= E[g_i(w_n)] - \gamma\sqrt{var[g_i(w_n)]} \end{aligned} \tag{10}$$

where γ is chosen based on the desired confidence probability. For example, $\gamma = 2\sigma$ gives 95% confidence probability for the Gaussian distribution.

4.2. Sigtest for The Available Data

In the following we show how the proposed signature boundaries in (10) can be used for the unimodality test. Let $z = [z_1, z_2, \cdots, z_N]$ be sorted absolute values of available data y, which its unimodality is unknown. Using signatures in (2) and (5), we can define our signature tests (Sigtest) as following:

$$Sigtest_1 \quad : \quad g_1(z_n) = z_n \tag{11}$$

$$Sigtest_2 \quad : \quad g_2(z_n) = \frac{1}{n}\sum_{j=1}^{n} z_j \tag{12}$$

To test the unimodality of z, $g_i(z_n)$ should be compared with the probabilistic bounds (based on our defined signatures, here i can be 1 or 2):

$$c_n = \begin{cases} 0, & L(n) < g_i(z_n) < U(n) \\ 1, & otherwise \end{cases} \tag{13}$$

$$C = \frac{1}{N}\sum_{n=1}^{N} c_n \tag{14}$$

where c_n shows any mismatch between the bounds and the signature at index n, and C is the total counting index. Consequently, the test chooses one of the hypotheses in Section 3 based on the following comparison:

$$C \quad \underset{H_0}{\overset{H_1}{\gtrless}} \quad T \tag{15}$$

where T is the threshold for a chosen confidence probability. Algorithm 1 demonstrates steps of the Sigtest. The behavior of Sigtest for 95% confidence probability (T=0.4) on synthetic clusters is shown in Fig.3. The left figures are clusters (a single and two overlapped clusters). The right figures are the behavior of the associated Sigtest. Bounds of the unimodal distribution ($U(n)$ and $L(n)$) are in blue dashed lines, while the test data z is the red line. The Sigtest of the first cluster lies completely inside the boundaries ($C = 0$), while for both (c) and (e) Sigtest is out of the boundary test which results in large values for C (0.95 and 0.99). Therefore, the method splits clusters in (c) and (e).

Algorithm 1 Unimodality Signature Test

```
Input: input samples $x = \{x_i\}_{i=1}^{N}$, $x_i \in R^d$, threshold $T$.
Output: result of the splitting test, split = 0 or 1.
$C \leftarrow 0$
$y \leftarrow f(x)$
$g_1(z_n) \leftarrow sort(abs(normalize(y)))$
$g_2(z_n) \leftarrow cumsum(g_1(z_n))$
$compute\ U(n)\ and\ L(n)\ from\ \ (10)$
for $j = 1$ to $N$ do
    if $g_i(z_j) > U(j) \quad or \quad g_i(z_j) < L(j)$ then
        $C \leftarrow C + 1$
    end if
end for
if $C > T$ then
    $split \leftarrow 1$
else
    $split \leftarrow 0$
end if
```

4.3. Role of the Signatures

In this paper we have proposed two signatures. While the first signature $g_1(w_n)$ deals with unimodal Gaussian distributions, the second signature $g_2(w_n)$ can work with any distribution from the unimodal family due to the weak law of large numbers, as the summation in (5) converges to Gaussian distribution for a large length of n. Consequently, even if data samples are non-Gaussian, $g_2(w_n)$ behaves similar to Gaussian.

5. SIMULATION RESULTS

In the first set of simulations, AD, KS, dip test, and Sigtest were used to examine the unimodality of two overlapped

[1] According to the Cauchy-Schwarz inequality we have:

$$cov(w_k, w_l) \leq \sqrt{var[w_k]}\sqrt{var[w_l]} \tag{7}$$

consequently:

$$\begin{aligned} var[g_2(w_n)] &= \frac{1}{n^2}(\sum_{j=1}^{n} var[w_j] + \sum_{k \neq l \geq 1}^{n} cov(w_k, w_l) \\ &\leq \frac{n}{n^2} var[w_n] + \frac{2}{n^2}\frac{n(n-1)}{2} var[w_n] \end{aligned} \tag{8}$$

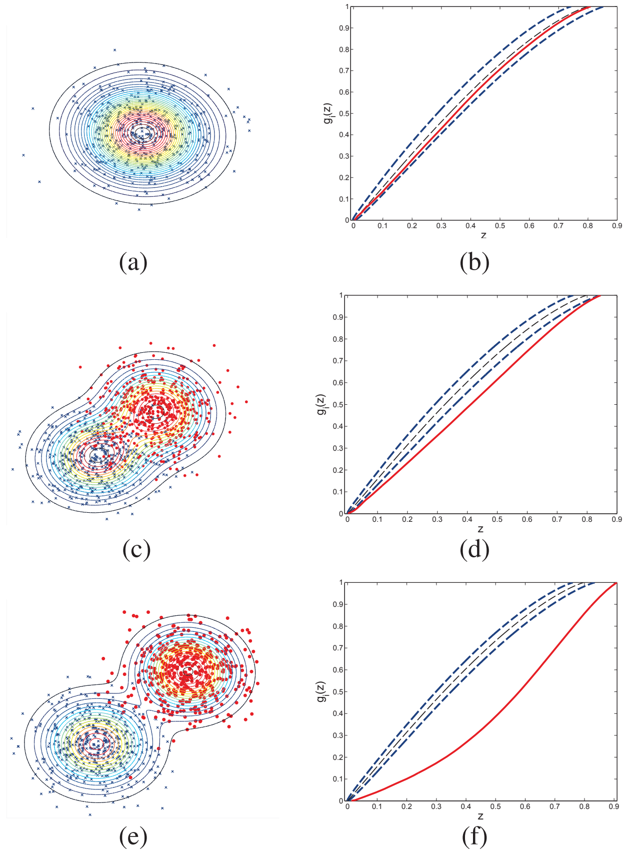

Fig. 3: Sigtest for single (a), and overlapped clusters (c) and (e). The two clusters (c) and (e) are separated with 2σ and 3σ respectively.

Gaussian clusters. Each cluster has 100 samples and its variance is $\sigma^2 = 1$. Table.1 shows the success rate of each statistic test when clusters are overlapped with different distances. The default significant levels are 0.0001 for AD, zero for Dip test, and 0.05 for KS. With 95% confidence probability in Sigtest, γ and T in (10) and (15) are 2 and 0.4 respectively. As the table shows the optimum methods (specially with more overlapping) are the Sigtests. In addition, the low computational complexity of the Sigtests resulted in much smaller computation time.

5.1. Application in Clustering

We denote dip-means and G-means when their splitting criteria are replaced with Sigtest as G-means$^+$ and dip-means$^+$. The comparison results on benchmark datasets are presented in Table 2. The quality of clustering is examined by Variation of Information (VI) [13] and Adjusted Rand Index (ARI) [14], where smaller VI and larger ARI are desired. Here, m^* is the correct number of clusters, and d is dimension of the data. Results are given for an average over 20 simulations. As the table shows G-means$^+$ and dip-means$^+$ consistently perform better than dip-means and G-means.

Table 1: Success rates of statistic tests for detecting two overlapped clusters with different central distances (averaged over 100 runs).

	Distance between center of clusters					
Tests	2σ	2.25σ	2.5σ	2.8σ	3σ	Average time (s)
$Sigtest_2(\%)$	56	93	99	100	100	1.96×10^{-4}
$Sigtest_1(\%)$	69	97	100	100	100	1.75×10^{-4}
AD(%)	29	76	97	100	100	3.96×10^{-4}
KS (%)	10	37	74	95	100	30×10^{-4}
dip (%)	3	8	21	82	94	2197×10^{-4}

Table 2: Comparison between G-means, dip-means and their improved version.

Data set	G-means	G-means$^+$	dip-means	dip-means $^+$
Iris	4.5±0.50	3±0	2±0	2.6±0.49
$(m^* = 3, d = 4)$				
VI	0.84±0.11	0.68±0.13	0.64±0.11	0.60±2.29
ARI	0.53±0.07	0.58±0.14	0.53±4.48	0.56±0.11
Optical digits	25.8±3.34	14.6±1.51	1±0	6.2±1.64
$(m^* = 10, d = 64)$				
VI	1.31±0.08	1.14±0.10	2.3025±0	1.76±0.09
ARI	0.57±0.03	0.66±0.04	0±0	0.35±0.05
Leukemia	4±0	3±0	1.75±0.44	3.1±0.41
$(m^* = 3, d = 39)$				
VI	0.49±0.00	0.30±0	0.84±0.14	0.67±0.18
ARI	0.77±0.00	0.88±0	0.39±0.23	0.59±0.14
Seed	4±0	2±0.72	1±0	3±0
$(m^* = 3, d = 7)$				
VI	0.87±0.00	0.84±0.15	1.0986±0	0.66±0
ARI	0.41±0.26	0.61±0.00	0±0	0.71±0
Pendigits	77.2±2.49	24.4±3.20	7±0	10.2±0.44
$(m^* = 10, d = 16)$				
VI	2.01±0.03	1.38±0.01	1.5866±0	1.4013±0.00
ARI	0.27±0.01	0.51±0.01	0.34±0	0.57±0.00

6. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced the idea of using signature test (Sigtest) for cluster splitting criterion. The two proposed signatures can compress the data based on its statistics and represent it in a space with smaller variation. The advantage of the proposed Sigtest compared to similar methods is in its robustness for recognizing the overlapped clusters, while its complexity is much less than the compared methods. The simulation results shows that replacing the existing splitting tests with Sigtest in hierarchical clustering improves the accuracy of estimated number of clusters as well as clustering quality. As future work, more signatures can be proposed for a general unimodal distribution or for a specific distribution in splitting criterion.

7. REFERENCES

[1] R. Xu, D. Wunsch, et al., "Survey of clustering algorithms," *Neural Networks, IEEE Transactions on*, vol. 16, no. 3, pp. 645–678, 2005.

[2] M.T. Chiang and B. Mirkin, "Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads," 2010, vol. 27, pp. 3–40, Springer.

[3] M. A. Stephens, "Edf statistics for goodness of fit and some comparisons," *Journal of the American statistical Association*, vol. 69, no. 347, pp. 730–737, 1974.

[4] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[5] J. A. Hartigan and P.M. Hartigan, "The dip test of unimodality," *The Annals of Statistics*, pp. 70–84, 1985.

[6] D. Pelleg and A. W. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, USA, 2000, ICML '00, pp. 727–734, Morgan Kaufmann Publishers Inc.

[7] M. Shahbaba and S. Beheshti, "Improving x-means clustering with mndl," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*. IEEE, 2012, pp. 1298–1302.

[8] G. Hamerly and C. Elkan, "Learning the K in K-Means," in *Neural Information Processing Systems*, 2003, vol. 17.

[9] X. Hu and L. Xu, "A comparative study of several cluster number selection criteria," in *Intelligent Data Engineering and Automated Learning*, pp. 195–202. Springer, 2003.

[10] Y. Feng and G. Hamerly, "PG-means: learning the number of clusters in data," in *Advances in Neural Information Processing Systems 19*, 2007, pp. 393–400.

[11] A. Kalogeratos and A. Likas, "Dip-means: an incremental clustering method for estimating the number of clusters," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 2402–2410.

[12] S. Beheshti, M. Hashemi, X.P. Zhang, and N. Nikvand, "Noise invalidation denoising," *Signal Processing, IEEE Transactions on*, vol. 58, no. 12, pp. 6007–6016, 2010.

[13] M. Meilă, "Comparing clusteringsan information based distance," *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007.

[14] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.