INFORMATICA, 2009, Vol. 20, No. 2, 187–202

On a Minimal Spanning Tree Approach in the Cluster Validation Problem

Zeev BARZILY [1], Zeev VOLKOVICH [1], Başak AKTEKE-ÖZTÜRK [2], Gerhard-Wilhelm WEBER [2,3]

[1] ORT Braude College of Engineering, 21982 Karmiel, Israel
[2] Institute of Applied Mathematics, Middle East Technical University, 06531 Ankara, Turkey
[3] Faculty of Economics, Business and Law, University Siegen, Hölderlinstrasse 3, 57076 Germany
e-mail: zbarzily@braude.ac.il, vlvolkov@braude.ac.il, bozturk@metu.edu.tr, gweber@metu.edu.tr

Received: August 2008; accepted: December 2008

Abstract. In this paper, a method for the study of cluster stability is purposed. We draw pairs of samples from the data, according to two sampling distributions. The first distribution corresponds to the high density zones of data-elements distribution. Thus it is associated with the clusters cores. The second one, associated with the cluster margins, is related to the low density zones. The samples are clustered and the two obtained partitions are compared. The partitions are considered to be consistent if the obtained clusters are similar. The resemblance is measured by the total number of edges, in the clusters minimal spanning trees, connecting points from different samples. We use the Friedman and Rafsky two sample test statistic. Under the homogeneity hypothesis, this statistic is normally distributed. Thus, it can be expected that the true number of clusters corresponds to the statistic empirical distribution which is closest to normal. Numerical experiments demonstrate the ability of the approach to detect the true number of clusters.

Keywords: clustering, cluster validation, minimal spanning tree, two sample test.

1. Introduction

Clustering deals with the partitioning of a data set to groups of similar items. In this paper, we address the cluster validation problem. Namely, we propose a new method intended to determine the true number of clusters. Iterative clustering procedures are frequently carried out in two phases: a partitioning phase, where the data set is divided into groups, and a validation phase, where the quality of the obtained partition is evaluated. Stability based approaches, for the cluster validation problem, evaluate the partitions variability under repeated applications of a clustering algorithm. Low variability is understood as high consistency in the results obtained (see, for example, Cheng *et al.*, 1996). Therefore, the number of clusters that maximizes cluster stability is accepted as an estimate for the “true” number of clusters in the data set.

We adopt the approach that the clusters correspond to modes of the probability density function of the data elements. Apparently, Wishart (1969) firstly promoted seeking for modes as an indication of the cluster structure. He suggested that clustering methods must be able to expose distinct data modes, independently of their shape and variance. Hartigan (1975), resting upon Wishart's idea, introduced the notion of high probability density clusters. The number of clusters has been recognized here as the amount of disjoint areas whose densities exceed a predefined value. An example consisting of three clusters in the plane with their cores is presented in Fig. 1. This simplified example exhibits a mixture of three two dimensional, same sized and spherical Gaussian distributions with the same standard deviation. Each component constitutes a cluster and its core is represented by a circle of radius 1.5 times the standard deviation. The clusters are centered at the distributions peak.

Our approach evaluates the goodness of a cluster by the likeness of clusters obtained from high and low densities zones. From Wishart's point of view, it is natural to foresee that, in the case of a stable partition, clusters defined by high density zones have to be similar to those defined by the whole data. We use probability metrics-distances, produced by two-sample test statistic, to measure the dissimilarity between clusters. The distances are calculated in clustered samples which are drawn from the source population according to two distributions. The first is constructed so that it represents the clusters high density cores and the second is constructed so that it represents the low density zones. Namely, we compare pairs of partitions and, a pair is considered to be consistent if the obtained clusters match. In this work, this matching is estimated via a Minimal Spanning Tree (MST), constructed for each one of the clusters, and followed by a computation of the number of edges connecting points from the two samples. Indeed, we employ here the Friedman–Rafsky's MST two sample test statistic (see, Friedman and Rafsky, 1979). Apparently, applying MST for clustering problems has been firstly proposed in Zahn (1971). Class discovery and feature selection via MST, for microarray gene expression data, have been considered in Varma and Simon (2004). Smith and Jain (1984) offered a method for

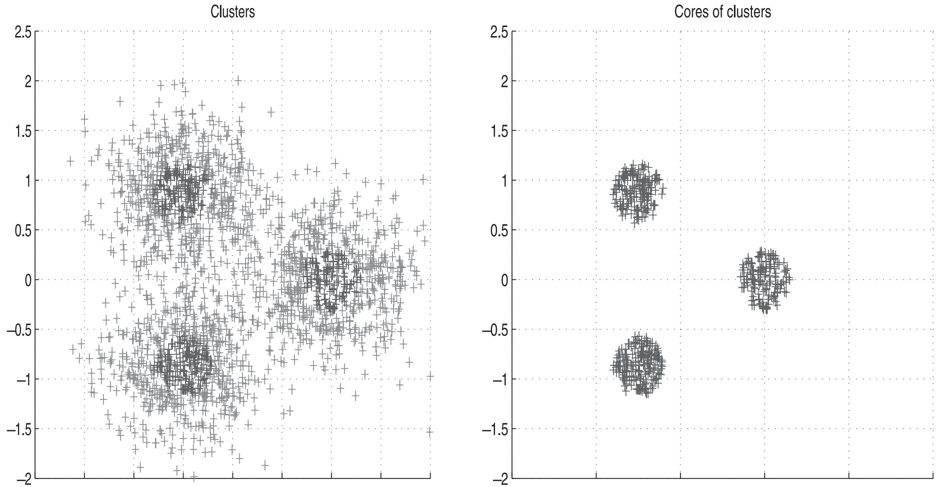

Fig. 1. An example of clusters and their cores.

testing uniformity, in multidimensional data, based on the Friedman–Rafsky's MST test. This approach has been extended in Jain *et al.* (2002).

2. Cluster Stability Model

Let us consider a finite subset $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$, of the d-dimensional Euclidean space R^d, drawn according to an underlying probability density function f_X. For a given set $S \subset X$, a partition $\Pi_k(S)$ of S is defined as a collection of subsets of S.

$$\Pi_k(S) = \{\pi_1(S), \pi_2(S), \ldots, \pi_k(S)\},$$

which satisfies the following conditions:

$$\bigcup_{j=1}^{k} \pi_j(S) = S$$

and

$$\pi_i(S) \cap \pi_j(S) = \phi, \quad i \neq j.$$

The elements of $\Pi_k(S)$ are called *clusters*. As usual, partitions are obtained by means of a clustering algorithm. We assume that such an algorithm Δ is available and that its input includes the data to be clustered and the desired number of clusters k. A partition of the data is the output of the algorithm. Within our model we apply the algorithm Δ to a set which is the union of two samples. The first sample is drawn, from the dataset X, according to the "cores distribution" having the probability density function

$$h_X(x) = \exp\big(af_X(x)\big)/c(a) \quad (x \in X),$$

and the second sample is drawn, from the dataset X, according to the "cores surroundings distribution" having the probability density function

$$g_X(x) = \exp\big(-af_X(x)\big)/c(-a) \quad (x \in X),$$

where

$$c(a) = \sum_{x \in X} \exp\big(af_X(x)\big),$$

$a > 0$ is a constant.

Let us suppose that we have a family of partitions Π_k $(k = 2, 3, \ldots, \tilde{k})$, where $\tilde{k}$ is some predefined number representing the maximal considered number of clusters. The proposed approach conjectures that, within partitions built for each possible number of clusters, the most stable one appears when the elements of the two samples are maximally close inside the clusters. In other words, within the clusters, the differences between the samples inside the clusters are expected to be minimal if the number of clusters is chosen correctly. As was mentioned early, we measure this dissimilarity by means of a two sample test statistic. Two sample tests are intended to test the null hypothesis which suggests that the elements of two considered samples have been drawn from the same distribution. Obviously, this is not the case here. However, we believe that, inside stable well-defined clusters, the elements of the two samples are mingled as if they were selected from close distributions.

Let $\boldsymbol{Y}^{(1)}$: $\{y_i^{(1)}, i = 1, \dots, t\}$ and $\boldsymbol{Y}^{(2)}$: $\{y_i^{(2)}, i = 1, \dots, s\}$ be two samples of independent random vectors having the probability laws F and G, respectively. The classical two-sample problem tests the null hypothesis

$$H_0: \quad F(x) = G(x)$$

against the general alternative

$$H_1: \quad F(x) \neq G(x),$$

in the case when the distributions F and G are unknown. The classical Kolmogorov–Smirnov test, the Cram'er-von Mises test, the Friedman's nonparametric ANOVA test and the Wald-Wolfowitz test must be reminded as the classical univariate procedures for this purpose. Many multivariate tests can be found in the literature (see, for example, Conover *et al.*, 1981; Duran, 1976; Friedman and Rafsky, 1979). Kernel based two sample tests were applied to this problem in the paper of Volkovich *et al.* (2008).

In this work we use the *Friedman–Rafsky's MST two* sample test statistic (the MST statistic). It considers an MST created for the pooled sample

$$\boldsymbol{V}_{t,s} = \boldsymbol{Y}^{(1)} \cup \boldsymbol{Y}^{(2)}$$

and defines the statistic as the number of edges of the MST, which connect a sample $\boldsymbol{Y}^{(1)}$element to a sample $\boldsymbol{Y}^{(2)}$ element. If all distances between elements of V are distinct, the set is called nice. Recall, that in this case, there is only one MST that connects all points of V, such that sum of the lengths of the edges is minimal. An MST for the set can be built in $O(|V|^2)$ time, including distance calculations, using the well known Prim's, Kruskal's, Boruvka's or Dijkstra's algorithms (see, for example, Nesetril *et al.*, 2001).

Under the null hypothesis, the MST statistic is normally distributed. More specifically, the following result was obtained. Let us denote by $R_{t,s}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)})$ the value of the MST statistic calculated for two given samples $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{Y}^{(2)}$. Actually, Friedman and Rafsky introduced the statistic $R_{t,s}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) + 1$, which represents the amount of disjoint sub-trees that result from removing all edges connecting points from distinct samples.

Theorem 1 (see, Theorem 1; Henze and Penrose, 1999). *Let*

- $s \to \infty$ *and* $t \to \infty$ *such that* $t/(s+t) \to p \in (0,1)$*;*
- *for each* t *and* s *the set* $\boldsymbol{V}_{t,s}$ *is nice,*

then the value

$$Z_{t,s} = \frac{1}{\sqrt{t+s}}\Big(R_{t,s}\big(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}\big) - \frac{2ts}{t+s}\Big),$$

convergences in distribution to the normal distribution with expectation 0 *and variance*

$$\sigma^2 = r\big(r + C_d(1-2r)\big),$$

where $r = 2p(1-p)$ *and* C_d *is a constant depending only on the space dimensionality.*

Now, we apply this Theorem 1 to characterize cluster stability. Theorem 1 implies that the problem can be reduced to estimation of distances from the empirical distribution of the statistic $R_{t,s}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)})$ to a normal one for various values of the possible number of clusters. The true number of clusters is assumed to correspond to the minimal distance. To motivate our approach let us consider two illustrative examples: An example of "a good cluster" is presented in Fig. 2. The number of edges connecting points from different samples is relatively large here – 25 out of 59. Indeed, the samples are sufficiently mingled so the clusters defined by them actually coincide.

The following example exhibits a case of two well separated clusters defined by samples. Here, the samples are artificially united by a single edge (Fig. 3).

The samples are, indeed, separated here and identify two different clusters. The MST statistic gets here its minimal value 1. Thus, the statistic value can indicate the cluster quality.

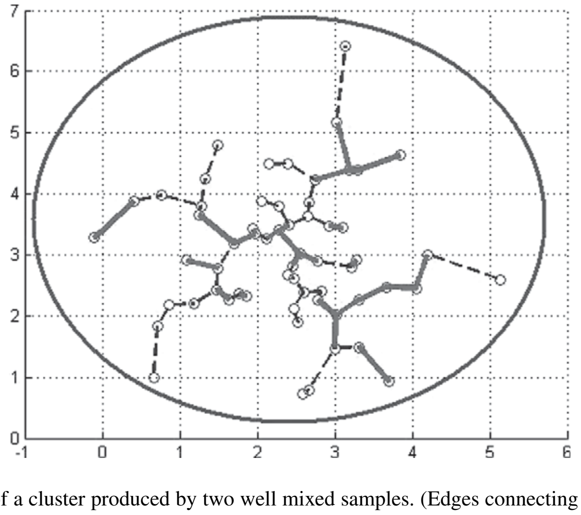

Fig. 2. An example of a cluster produced by two well mixed samples. (Edges connecting points from different samples are marked by solid lines.)

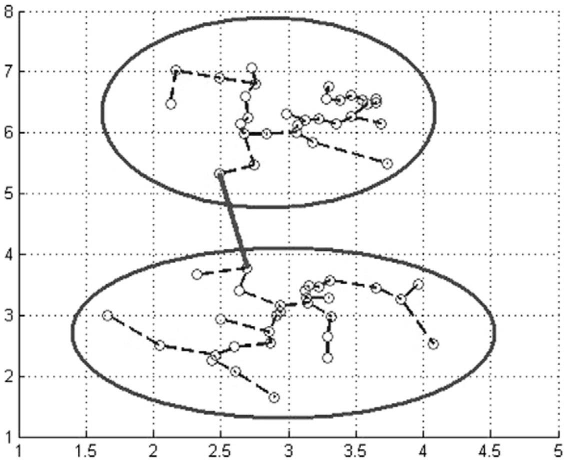

Fig. 3. An example of clusters produced by separated samples. (The edge connecting points from different samples is marked by a solid line.)

For a given partition $\Pi_k(\boldsymbol{X})$ an underlying probability density function f_X can be represented by the form

$$f_{\boldsymbol{X}} = \sum_{i=1}^{k} p_i f_i,$$

where p_i, $i = 1, \ldots, k$ are the clusters' probabilities and f_i, $i = 1, \ldots, k$ are the clusters' densities. Let $R_{t,s}^{(i)}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)})$ be the value of the MST statistic calculated within the cluster number i. We introduce

$$R_{t,s,k}^{(0)}\big(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}\big) = \min_{i=1,\ldots,k} \left\{ \sqrt{\frac{k}{2(t+s)}} \left(R_{t,s}^{(i)}\big(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}\big) - \frac{2ts}{k(t+s)} \right) \right\}.$$

While comparing a partition to some other (base) partition, it must be noted that a cluster's identity is usually represented by its meaningless label. The same cluster in two partitions, obtained by different applications of a clustering process, can be differently labeled. Thus, once a convergence of partitions $\{\Pi_{k,m}\}$, $m = 1, 2, \ldots$, is considered, we have to assume that the labeling of their elements are changed, on each step, by a suitable sequence of permutations $\psi_m(i)$ of the label set $\{1, \ldots, k\}$.

Theorem 2. *For a given number of clusters k:*

1. *There exists a partition $\Pi_k^{(0)}$, of the set $\boldsymbol{X}$, such that, for each sequence of samples $\{V_m\}$, the sequence of partitions*

$$\Pi_{k,m}(V_m) = \{\pi_{1,m}, \ldots, \pi_{k,m}\} = \Delta(V_m, k)$$

converges to $\Pi_k^{(0)}$*, for a sequence* $\psi_m(i)$ *of permutations of the label set* $\{1,\dots,k\}$*. The convergence is in the sense that*

$$\lim_{|V_m|\to\infty}\sum_{i=1}^{k}\left|\pi_{\psi_m(i),m}\oplus\pi_i^{(0)}\right| = 0.$$

Here $\oplus$ *denotes the symbol XOR (the symmetric difference operation) and* $|A|$ *is the cardinality of a set A.*

2. *The partition* $\Pi_k^{(0)}$ *satisfies*

$$p_1 = p_2 = \ldots = p_k.$$

Then for two sequences of samples $\{\boldsymbol{Y}_m^{(1)}\}, \{\boldsymbol{Y}_m^{(2)}\}$*, having equal size* $t_m\to\infty$*, such that* $\boldsymbol{V}_m = \boldsymbol{Y}_m^{(1)}\cup\boldsymbol{Y}_m^{(2)}$ *is nice for each* m*, the following random variable converges in distribution,*

$$R^{(0)}_{t_m,t_m,k}\big(\boldsymbol{Y}_m^{(1)},\boldsymbol{Y}_m^{(2)}\big)\to G_k,$$

where G_k *is a random variable representing the minimal value of* k *independent identically distributed standard normal variables.*

Proof. We consider a sequence of equal sized pairs of samples $|\boldsymbol{Y}_m^{(1)}|=|\boldsymbol{Y}_m^{(2)}| = t_m$ with nice union sets $\boldsymbol{V}_m$. Let $\{\pi_{i,m}\}, m = 1,2,\dots, i = 1,\dots,k$ be a sequence of clusters, defined by $\boldsymbol{V}_m$. Resting upon the MST, constructed for $\pi_{i,m}$, we introduce the variables

$T_m^{(i)}$ – the number of edges connecting points from different samples belonging to $\pi_i^{(0)}$;

$S_m^{(i)}$ – the number of edges connecting points from different samples where at least one of the points does not belong to $\pi_i^{(0)}$.

Thus, the value of the MST statistic, built within $\pi_{\psi_m(i),m}$, equals

$$M_m^{(i)} = T_m^{(i)} + S_m^{(i)}$$

and

$$\Big(T_m^{(i)} - R^{(i)}_{t_m,t_m,k}\big(\boldsymbol{Y}_m^{(1)},\boldsymbol{Y}_m^{(2)}\big)\Big)\to 0$$

in distribution. Moreover, due to the conditions mentioned above

$$\lim_{m\to\infty}\frac{k|\boldsymbol{Y}_m^{(1)}\cap\pi_{\psi_n(i),m}|}{t_m} = \lim_{m\to\infty}\frac{k|\boldsymbol{Y}_m^{(2)}\cap\pi_{\psi_n(i),m}|}{t_m} = 1,$$

for all $i = 1,\dots,k$ and a sequence $\psi_m(i)$ of permutations of the label set $\{1,\dots,k\}$. The Theorem 2 assertion is now obtained by substituting the last equalities in the statement of Theorem 1.

We would like to note that a convergence condition presented in the Condition 1 of the Theorem 2 is, indeed, a partial case of a general definition related to the convergence of clustering algorithms (see, for example, Von Luxburg and Ben-David, 2005, Section 5.2). For the k-means approach such a property has been discussed in Pollard (1981) and Lember (2003). Condition 2 is sufficientlywidespread in cluster analysis. For instance, Celeux and Govaert (1992) showed that the k-means approach is a specific case of the common EM approach, which assumes in the Gaussian mixture model with equal sized clusters having the same covariance matrix $\Gamma_i = \sigma^2 I$, $i = 1, \dots, k$, where I is the identity matrix having the order d and σ^2 is an unknown parameter, i.e., the k-means algorithm essentially applies spherical, same sized clusters.

The presence of clusters having the same size and geometrical structure could be rarely guaranteed for real datasets. Nevertheless, these requirements are common. Consequently, the proposed model can be considered, for the distribution of the values $R^{(0)}_{t,t,k}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)})$, as a theoretical etalon characterizing a stable partition situation.

As was noted earlier, a sequence $\psi_m(i)$ appears in the theorem's statement because a cluster can be differently marked in rerunning of the clustering algorithm.

3. The Algorithm

According to our model the MST test is iterated many times, for each potential numbers of clusters, to yield an empirical MST test statistic distribution. The distribution, which is closest to normal, is assumed to correspond to the true number of clusters. Let us draw, without replacement, M pairs of samples $S_1^{(m)}$ and $S_2^{(m)}$ ($m = 1, \dots, M$), according to the densities h_X and g_X, respectively. We introduce the set

$$S^{(m)} = S_1^{(m)} \bigcup S_2^{(m)}$$

and the partition

$$\Pi_k^{(m)} = \Delta(S^{(m)}, k).$$

Hence, we consider

$$S_{1l}^{(m)} = S_1^{(m)} \cap \pi_l^{(m)}(S^{(m)}), \quad S_{2l}^{(m)} = S_2^{(m)} \cap \pi_l^{(m)}(S^{(m)})$$

as subsets of elements from $S_1^{(m)}$ and $S_2^{(m)}$, respectively belonging to the cluster $\pi_l^{(m)}(S^{(m)})$. We measure dissimilarities between these sets, within $\pi_l^{(m)}(S^{(m)})$, with the help of the MST statistic. Denote by $r_l^{(m)}$ ($l = 1, \dots, k$) the values of this statistic, calculated inside $\pi_l^{(m)}(S^{(m)})$, as the number of edges which connect elements of $S_{1l}^{(m)}$ to elements of $S_{2l}^{(m)}$. Thus, for each $k = 2, \dots, \tilde{k}$ we obtain an array of k-tuples having size M. We measure the distance, between the samples of pair number m, by

$$R^{(m)} = \min\left\{ r_l^{(m)} \mid l = 1, 2, \dots, k \right\}.$$

Thus, under the assumptions presented in Theorem 2, the random variable$R^{(m)}$ is approximately distributed as the minimal value of k i.i.d. random variables, distributed according to a normal distribution $N(\mu, \sigma^2)$ if k is chosen correctly. In order to estimate the mean μ and the variance σ^2 we calculate the average value of $r_l^{(m)}$ $(l = 1, \dots, k)$:

$$T^{(m)} = \text{Average}\big\{r_l^{(m)} \mid l = 1, 2, \dots, k\big\}.$$

From the M averages $T^{(m)}$ $(m = 1, 2, \dots, M)$, their $\{\text{mean}(T)\}$ and their variance $\{\text{var}(T)\}$ are obtained. For this end we substitute:

$$\mu = \text{mean}(T), \quad \sigma^2 = \text{var}(T).$$

As a result, we conclude, that the estimate of the "true" number of clusters maintains a distribution of the values $R^{(m)}$ which is the maximally close to the distribution of the minimum of k independent normal variables with the chosen μ and σ^2. For this purpose we simulate M values

$$W^{(m)} = \text{Min}\big\{Z_l \mid l = 1, \dots, k\big\},$$

where Z_l(l $=$1, $\dots, k$) are the appropriate normal i.i.d. variables. Note, it is well known that the distribution function $P_{W^{(m)}}(x)$ of $W^{(m)}$ can be expressed as

$$P_{W^{(m)}}(x) = 1 - \big(1 - P_{(\mu,\sigma^2)}(x)\big)^k,$$

where $P_{(\mu,\sigma^2)}(x)$ is the distribution function of $N(\mu, \sigma^2)$. These two possibilities often provide similar outcomes however; we prefer to compare two empirical distributions since approximations of $P_{(\mu,\sigma^2)}(x)$ can yield additional calculations instabilities. Therefore, we apply one-dimensional two sample test statistic to assess the distance between the empirical distributions of $R^{(m)}$ and $Y^{(m)}$. This can be done by means of the famous Kolmogorov–Smirnov distance, often called the K-S distance. It is applied to settle on whether two underlying one-dimensional probability distributions are different, or whether an underlying probability distribution fits a hypothesized distribution.

Remarks

- We indirectly assume that the sets $S_{1l}^{(m)}$ and $S_{2l}^{(m)}$ are permanently nice. This suggestion holds for almost all real datasets.
- The idea to explore differences between the high and low density regions by the *Friedman–Rafsky's MST* statistic is presented in (Smith and Jain, 1984) and (Jain *et al.*, 2002). The goal was to point out an "inconsistent" edge whose length is significantly larger than the average length of the nearby edges. We apply the MST two sample test statistic in the "departure from normality" manner that allows attainment of more detailed results about the data cluster structure.

- The high and low density regions have been characterized in the papers by Smith and Jain (1984) and Jain *et al.* (2002) via the MST which appears to be computational costly for big datasets. We use a variant of the nearest neighbors approach for the underlying density estimation.

4. Experimental Results

In order to evaluate our methodology, we provide several numerical experiments on synthetic and real datasets. We demonstrate the performance of the proposed method by comparing the obtained results to the assumed “true” structure, on two simulated and three real datasets. Estimates of the core density, h, and the core surroundings density, g, are constructed by replacing $f(x)$ by its estimate, obtained via a variant of the k-nearest neighbors. It is set in the following way: for each point $x \in X$ we determine its k nearest neighbors $Y_k = \{y_1, y_2, \dots, y_k\}$. Here k is chosen as the minimal value between 100 and $|X|/2$ and a sphere radius is calculated by

$$R = \frac{1}{10k} \sum_{i=1}^{k} \|x - y_k\|.$$

An estimated value of $f(x)$ is found as

$$\tilde{f}(x) = \frac{\bigl|y \in Y_k \colon \|x - y_k\| < R\bigr|}{|X|}.$$

Note, the meaning of k here differs from its meaning in the rest of the paper.

We provide 10 trials for each experiment. The results are presented via the error-bar plots of the K-S distance mean within the trials. The sizes of the error bars are two standard deviations, calculated within the trials. The regular k-means algorithm is selected as the clustering algorithm. We choose $a = \ln(4)$ and the maximal tested number of clusters is 7.

4.1. *Artificial Datasets*

These datasets have been simulated as 4000 points drawn from a mixture of two dimensional, spherical Gaussian distributions with the same standard deviation. The first dataset consists of three components, centered at the coordinates $(\cos(\frac{2\pi i}{3}), \sin(\frac{2\pi i}{3})), i = 0, 1, 2$, and having standard deviations of 0.4. Two components include 1333 items and the last one includes 1334 items. Scatter plot of this data is presented in Fig. 4. Error-bar plot, for $M = 100$ and a sample size of 225, is presented in Fig. 5.

The second dataset has been constructed as a mixture of five equal sized components, centered at the points $(\cos(\frac{2\pi i}{5}), \sin(\frac{2\pi i}{5}))$, $i = 0, 1, 2, 3, 4$, and having standard deviations of 0.2 (Fig. 6). Error-bar plot, for $M = 300$ and a sample size of 700, is presented in Fig. 7.

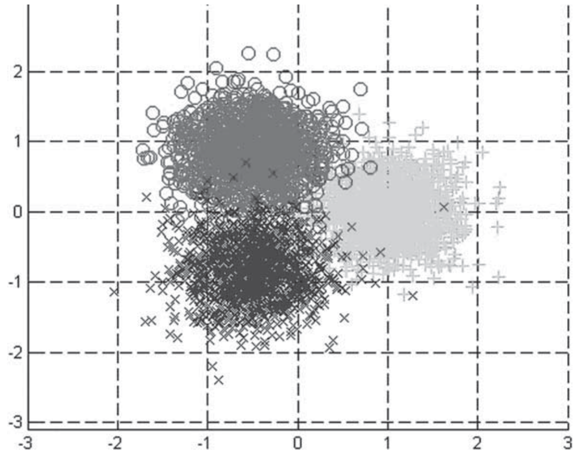

Fig. 4. Scatter-plot of the three components simulated data.

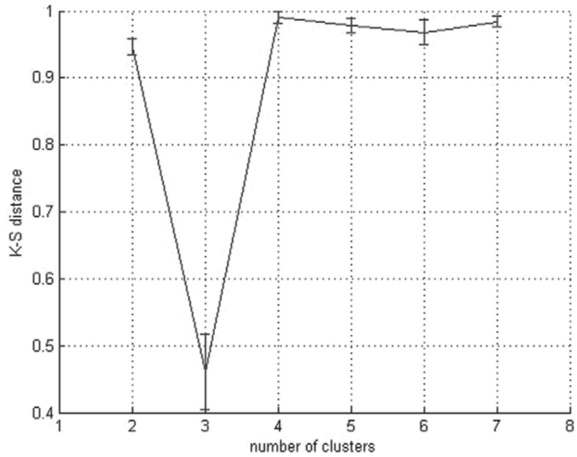

Fig. 5. Error-bar plot of the K-S distance of the three components simulated data.

4.2. *Three Text Collection Datasets*

The dataset has been chosen from `http://ftp.cs.cornell.edu/pub/smart/`.

It includes

- DC0–Medlars Collection (1033 medical abstracts);
- DC1–CISI Collection (1460 information science abstracts);
- DC2–Cranfield Collection (1400 aerodynamics abstracts).

This dataset has been analyzed in several works (see, for example, Kogan *et al.*, 2003 and Volkovich *et al.*, 2004). Following the well-known "bag of words" approach, 300 and 600 "best" terms are selected and the dataset is embedded into Euclidian spaces having dimensions identical to the terms quantity. Naturally, the dataset constructed by means of the 300 best terms can be considered as a noisy version of the one based on the 600 best

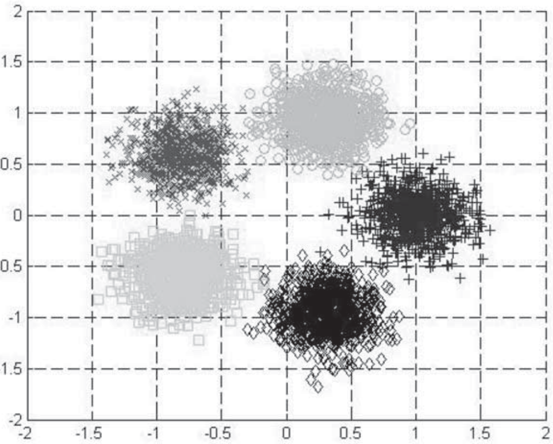

Fig. 6. Scatter-plot of the five components simulated data.

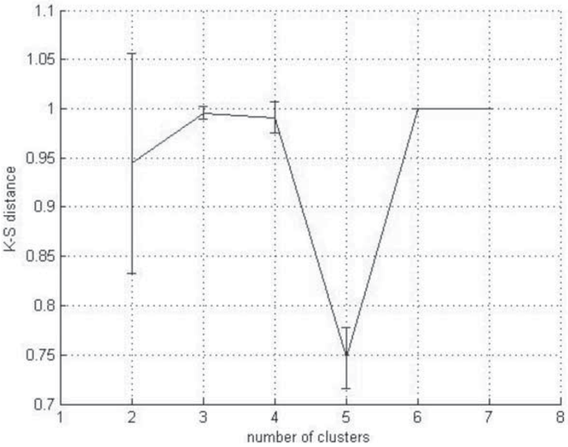

Fig. 7. Error-bar plot of the five components simulated data.

terms. A dimension reduction, in both cases, is carried out by the Principal Component Analysis. The dataset is identified to be well separated via two leading principal components. Therefore, we use this data representation in our experiments. Here we choose $M = 100$ and the sample size equals 300.

Another experiment has been provided in order to investigate the approach steadiness with this dataset constructed resting upon only 300 terms. Such a data can be viewed as a blurred version of the previous one. Here more unstable outcomes are expected. However, the true number of clusters has been, again, clearly detected.

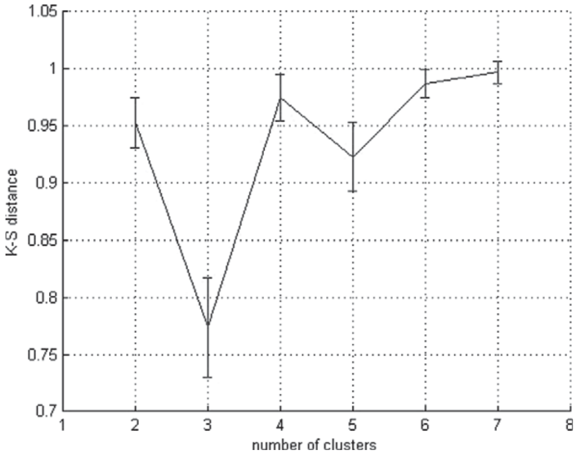

Fig. 8. Error-bar plot of the K-S distance for the three text collection dataset for 600 terms.

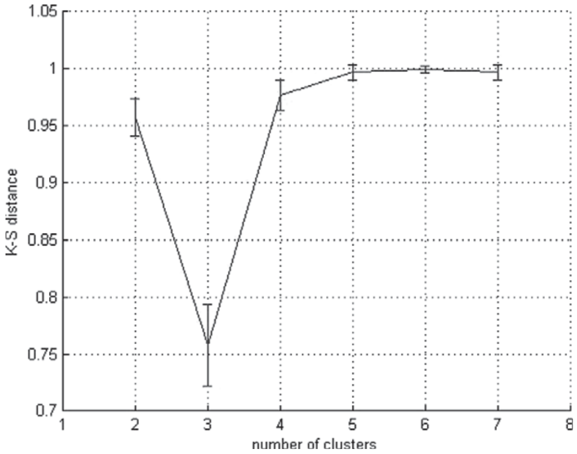

Fig. 9. Error-bar plot of the K-S distance for the three text collection dataset for 300 terms.

4.3. *The Iris Flower Dataset*

Additional considered dataset is the well known Iris Flower Dataset. It is available, for example, at `http://archive.ics.uci.edu/ml/datasets/Iris`. This dataset contains four-dimensional feature vectors for three distinct sets of flowers. There are 50 entries for each class. It is well known that one cluster is linearly separable from the others while the other two are not. This data set was analyzed in many papers. For instance in Roth *et al.* (2002) a two cluster structure was detected. However, our approach is capable to identify a three cluster structure. Here we choose $M = 200$ and the sample size equals 70.

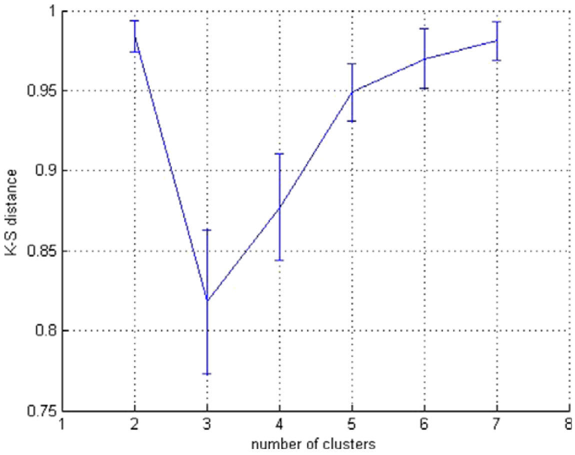

Fig. 10. Error-bar plot of the K-S distance for the Iris dataset.

5. Conclusions

We have developed a novel method for cluster validation based on the Minimal Spanning Tree two sample test. The method suggests to measure the clusters quality by means of the test statistic applied to pairs of clustered samples drawn from high and low density regions. A new approach, for the departure from normality estimation of an empirical statistic distribution, is offered.

All experiments performed detected the true number of clusters. The result of the three components dataset, even though some overlapping of the observed occurrences area exists, clearly detects the three clusters structure. The five components dataset contains well separated components. Thus, it is of no surprise that the correct number of clusters is obtained here. The analysis of the abstracts dataset is carried out twice – with 600 terms and with 300 terms. We expected better results in the first case but, it did not materialize. Apparently, 300 terms are sufficient for representing the abstracts variability.

In all cases addressed so far, due to the significance of the results, additions or subtractions of two standard deviations do not generate an overlap of K-S-distance value in the neighborhood of the true number of clusters.

The Iris Flower dataset is the most difficult to analyze. It is due to the fact that two clusters are not linearly separable. The true number of clusters is detected but the result of four clusters is near optimal. This fact may need a more specific analysis.

References

Celeux, G., and G. Govaert (1992). A classification EM algorithm and two stochastic versions. *Computational Statistics and Data Analysis*, **14**, 315–332.

Cheng, R., and G.W. Milligan (1996). Measuring the influence of individual data points in a cluster analysis. *Journal Classification*, **13**, 315–335.

Conover, W.J., M.E. Johnson and M.M. Johnson (1981). Comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, **23**, 351–361.

Duran, B.S. (1976). A survey of nonparametric tests for scale. *Communications in Statistics – Theory and Methods*, **5**, 1287–1312.

Friedman, J.H., and L.C. Rafsky (1979). Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, **7**, 697–717.

Hartigan, J.A. (1975). *Clustering Algorithms*. John Wiley, New York.

Jain, A.K., X. Xu, T.K. Ho and F. Xiao (2002). Uniformity testing using minimal spanning tree. In *Proc. ICPR*, Vol. 4. pp. 281–284.

Henze, N., and M. Penrose (1999). On the multivariate runs test. *Annals of Statistics*, **27**(1), 290–298.

Kogan, J., C. Nicholas and V. Volkovich (2003). Text mining with information – theoretical clustering. *Computing in Science and Engineering*, November/December, 52–59.

Lember, J. (2003). On minimizing sequences for k-centers. *Journal of Approximation Theory*, **120**, 20–35.

Von Luxburg, U., and S. Ben-David (2005). Towards a statistical theory of clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*.

Nesetril, J., E. Milkova and H. Nesetrilova (2001). Otakar boruvka on minimum spanning tree problem: Translation of both the 1926 papers, comments, history. *Discrete Mathematics*, **233**(1), 3–36.

Pollard, D., (1981). Strong consistency of k-Means clustering. *Annals of Statistics*, **9**(1), 135–140.

Roth, V., V. Lange, M. Braun and J. Buhmann (2002). A resampling approach to cluster validation. In *Proc. Intl. Conf. on Computational Statistics*. pp. 123–128.

Smith, S.P., and A.K. Jain (1984). Testing for uniformity in multidimensional data. *IEEE Transactions Pattern Anal. Machine Intelligence*, **6**(1), 73–80.

Varma, S., and R. Simon (2004). Iterative class discovery and feature selection using minimal spanning trees. *BMC Bioinformatics*, **5**, 126.

Volkovich, V., J. Kogan and C. Nicholas (2004). k-means initialization by sampling large datasets. In I. Dhillon and J. Kogan (Eds.), *Proceedings of the Workshop on Clustering High Dimensional Data and its Applications* (held in conjunction with *SDM 2004*). pp. 17–22.

Volkovich, Z., Z. Barzily and L. Morozensky (2008). A statistical model of cluster stability. *Pattern Recognition*, **41**(7), 2174–2188.

Wishart, D. (1969). Mode analysis: A generalization of nearest neighbors which reduces chaining effects. *Numerical Taxonomy*, 282–311.

Zahn, C. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, **20**, 68–86.

Z. Barzily, associate professor is a head of the Information System Engineering Program at Ort Braude College, Karmiel, Israel. He received his PhD in operations research in 1974 from the Technion, Israel Institute of Technology. Dr. Barzily has been doing research in stochastic modeling, and in cluster's stability evaluation.

Z. Volkovich, associate professor is a head of the Department of Software Engineering Ort Braude College, Karmiel, Israel and also affiliated with the Department of Mathematics and Statistics, University of Maryland at Baltimore County. He received his PhD in probability theory and mathematical statistics in 1982. Dr. Volkovich has been doing research in statistical pattern recognition, feature selection, text mining, stochastic models, and generalized convolutions. His recent research includes classification with unsupervised learning and clustering evaluation.

B. Akteke-Öztürk is a PhD student at the Scientific Computing Department of the Institute of Applied Mathematics, Middle East Technical University, Turkey. She received her MSc from the Institute of Applied Mathematics, Middle East Technical University, Turkey in 2005. Her research areas is nonsmooth optimization and data mining algorithms for the quality improvement problems.

G.-W. Weber, professor works at IAM of METU, Ankara, Turkey, in the Department of Financial Mathematics and Department of Scientific Computing. He is also a guest professor at the Faculty of Economics, Management and Law at the University of Siegen, Germany. He received his diploma and doctorate in mathematics and economics at RWTH Aachen. His research interest are lying in the areas of continuous optimization, financial mathematics, OR, optimal control, selected topics from discrete optimization, dynamical systems, statistical learning and computational statistics, inverse problems, complexity theory, computational biology, environmental protection, development and societal complexity.

Minimalaus jungimo medis klasterių skaičiui nustatyti

Zeev BARZILY, Zeev VOLKOVICH, Başak AKTEKE-ÖZTÜRK,
Gerhard-Wilhelm WEBER

Šiame straipsnyje pasiūlytas metodas klasterių stabilumui tikrinti. Iš duomenų aibės paimamos kelios poros poaibių pagal skirtingus pasiskirstymus, t.y. vieno poaibio elementų pasiskirstymas atitinka didelio tankio zoną, šie elementai susiję su klasterio centru; kito poaibio elementų pasiskirstymas atitinka mažo tankio zoną, šie elementai susiję su klasterio pakrasčiu. Elementai yra kelis kartus klasterizuojami nustačius įvairius klasterių skaičius. Teigiama, kad klasteriai yra pastovūs, jei gauti klasteriai yra panašūs. Panašumas yra įvertinamas pagal minimalaus jungimo medžio jungčių, jungiančių taškus iš skirtingų pasiskirstymų poaibių, skaičių. Yra naudojamas Friedmano–Rafskio dviejų elementų testo statistika. Pagal homogeniškumo hipotezę ši statistika yra pasiskirsčiusi pagal normalųjį dėsnį. Taigi, galima tikėtis, kad tikrasis klasterių skaičius atitiks šios statistikos empirinį pasiskirstymą, kuris yra arčiausiai normalinio. Skaitiniai eksperimentai parodo siūlomo metodo galimybę nustatyti tikrąjį klasterių skaičių.