A SINGLE-STEP CLUSTERING ALGORITHM BASED ON A NEW INFORMATION-THEORETIC SAMPLE ASSOCIATION METRIC DEFINITION

T. Temel[*]

Abstract: A single-step information-theoretic algorithm that is able to identify possible clusters in dataset is presented. The proposed algorithm consists in representation of data scatter in terms of similarity-based data point entropy and probability descriptions. By using these quantities, an information-theoretic association metric called mutual ambiguity between data points is defined, which then is to be employed in determining particular data points called cluster identifiers. For forming individual clusters corresponding to cluster identifiers determined as such, a cluster relevance rule is defined. Since cluster identifiers and associative cluster member data points can be identified without recursive or iterative search, the algorithm is single-step. The algorithm is tested and justified with experiments by using synthetic and anonymous real datasets. Simulation results demonstrate that the proposed algorithm also exhibits more reliable performance in statistical sense compared to major algorithms.

Key words: *clustering, clustering algorithms, information theory, mutual information, unsupervised learning*

Received: March 7, 2015
Revised and accepted: October 30, 2017

DOI: 10.14311/NNW.2017.27.027

1. Introduction

In a broader sense, a clustering algorithm is expected to perform a twofold operation: Given a dataset, estimating a model e.g. number of data point groups or clusters, with optimal fit and compactness, which is often abbreviated to by cluster validity [4], and partitioning the dataset in accordance with the model, e.g. number of clusters estimated as such. Major clustering algorithms rely on certain observable regularities in data scatter of data points/feature vectors to be statistically modeled. Hierarchical methods group data points by dividing/merging substructures recursively till no further connected group is achievable, which results in a dendogram. Although their time complexity is high, e.g. $O(N^3)$, time

[*]Turgay Temel; Bursa Technical University, Faculty of Natural Sciences, Architecture and Engineering, Mechatronics Engineering Dept. Osmangazi, Bursa, Turkey, E-mail: turgay.temel@btu.edu.tr, ttemel70@gmail.com

complexity of some agglomerative hierarchical algorithms with speed-up reduces to $O(N^2)$ where N is the number of data points [3, 13]. While most hierarchical algorithms do not require the *a priori* knowledge for the number of clusters they usually fail in case of overlapping densities and they are sensitive to outliers. Partitioning algorithms aim at obtaining a partition of dataset by assigning N data points in hyper-surfaces to k clusters based on an optimality criterion as a separation quality between clusters. They generally run iteratively to inspect variation of a distortion/compactness factor versus the *a priori* number for clusters which is varied within a prescribed range: For example, k-means [2], and mixture of varying densities/distributions [12], have usually time complexity of $O(kNI)$ where k is the *a priori* number of clusters being processed and I is the number of iterations which strictly depends on adopted termination rule. As a variant of partitioning methods, spectral algorithms are capable of clustering non-convex datasets with reduced dimensions by using k largest eigenvectors corresponding to Laplacian of the similarity matrix [6]. Their complexity is usually $O(N^{3/2} + kNI)$ or higher while employing faster partitioning methods, such as k-means in initialization and/or post processing steps. In density-based clustering, clusters are obtained as regions where data points are densely located: In DBSCAN [5], similar to linkage-based hierarchical clustering, those data points which are within a distance of presumed threshold form clusters subject to a minimum number of data points. In mean-shift clustering, each data point is moved to the possibly densest neighborhood based on estimated maxima of kernel density being sought [17]. A data point that resides at a (local) maximum of density after all data have moved is regarded as a centroid/mean. Both methods do not require the *a priori* knowledge of the number of clusters, and impose no constraint on the shape of the clusters. Since mean-shift algorithm highly relies on estimating the neighboring data points to which mean vectors are to be shifted at each successive step, it is computationally expensive with $O(N^2I)$ where I was cited previously compared to DBSCAN which usually has complexity $O(N \log N)$.

Information-theoretic notions, e.g. entropy and mutual-information, are invariant to data representation and capable of capturing higher-order statistics [14, 16]. Owing to these advantages and well-defined representation of data scatter properties in terms of entropy and mutual information as an association rule these quantities have been applied in clustering studies, e.g. kernel-based hierarchical clustering with use of optimized quadratic mutual-information [1], and clustering algorithm based on Renyi's entropy [10]. A recent study in [15] proposes a method to estimate the number of clusters in single-step by identifying data points at possible cluster boundaries in a dataset based on information-theoretic sample (or interchangeably referring to data point with presumed density profile) entropy and probability descriptions. To our knowledge, there is no algorithm that yields both the number of clusters and forms the respective clusters in single-step or one-pass.

In this study, a single-step clustering algorithm is presented with use of a new information-theoretic association measure that exploits a quantity called mutual ambiguity as an extension to information-theoretic notions given in [15]. It is illustrated that the new algorithm successfully extracts even non-convex clusters while preserving their actual shapes most major algorithms fail in. Simulations with a real dataset for new algorithm and some major counterparts reveal that

the new method statistically outperforms the counterparts in terms of successful identification of clusters and sample classification with much lower time complexity.

2. Similarity-based sample density and entropy and cluster analysis

A definition of similarity, s_{ij}, between D-dimensional data points $\mathbf{x}_i$ and $\mathbf{x}_j$ with $\mathbf{x}_i \neq \mathbf{x}_j$ is commonly given by

$$s_{ij} = e^{-\beta d_{ij}^2}, \tag{1}$$

where $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$ is the Euclidean distance (metric) between them. The parameter β is the kernel size or resolution coefficient and it is usually taken unity. Given a dataset of N data points, we define a similarity-based (experimental) sample probabilistic distribution (SBD) for which data point $\mathbf{x}_i$ is at the center as

$$p_i = \sum_{\forall j \neq i} s_{ij}, \tag{2}$$

where summation is taken over all $\mathbf{x}_j$ and $\mathbf{x}_j \neq \mathbf{x}_i$. It is possible to express differential variation in similarity as

$$\partial s_{ij} = -2\beta_i s_{ij}(\mathbf{x}_i - \mathbf{x}_j) \bullet \partial(\mathbf{x}_i - \mathbf{x}_j) = \nabla s_{ij} \bullet \partial(\mathbf{x}_i - \mathbf{x}_j), \tag{3}$$

where ∇s_{ij} refers to the gradient vector of similarity between $\mathbf{x}_i$ and $\mathbf{x}_i$ with respect to difference $\mathbf{x}_i - \mathbf{x}_j$ and '$\bullet$' is the dot-product. Density or mode seeking algorithms, e.g. mean-shift clustering, are based on identifying a data point $\mathbf{x}_i$ that meets

$$\sum_{\forall j \neq i} \nabla s_{ij} \bullet \partial(\mathbf{x}_i - \mathbf{x}_j) = 0 \tag{4}$$

in terms of local optima to be sought. It is seen that as other data points are brought closer to $\mathbf{x}_i$ respective SBD reaches a local maximum, which suggests a representation for data scatter of neighborhood. However, SBD is not suitable for characterizing data scatter in uncertain regions, e.g. overlapping or superposition of similarities, or uniformly distributed regions of data points. Similar to stochastic entropy definition, an experimental counterpart can be suggested, which we call similarity-based sample entropy (SBE) given by

$$H_i = H(\mathbf{x}_i) = -\sum_{\forall j \neq i} s_{ij} \log s_{ij} \tag{5}$$

as a measure of data scatter around $\mathbf{x}_i$. From above descriptions, it is straightforward to see close resemblance between information-theoretic probabilistic, i.e. stochastic entropy and its experimental similarity-based counterpart: Given a particular data point $\mathbf{x}_i$ its SBE decreases toward a minimum as other data points are moved either very close to or far from it. It reaches a maximum as other data points are positioned at a distance irregularly where it is difficult to assert on proximities.

Above quantities have been used in [15] to identify possible data regions for cluster availability within a given dataset in single-step. It employs a function called cluster-boundary indicator (CBI) given by $G = e^{-\varphi}$ where $\varphi = |p - \lambda H|^2$ with $\lambda > 0$. A data point $\mathbf{x}*$ which satisfies $\partial^2 G = 0$ for a maximum of G close to 1 refers to availability of a cluster shown as the shaded area in Fig. 1. A respective cluster is formed with data points which satisfy $|p - \lambda H| < 1$. Data point $\mathbf{x}_c$ is the mean or centroid of the cluster formed as such. The scaling factor λ is determined such that $\sum_{\forall i} (p_i - \lambda H_i)^2$ is minimum, which yields $\lambda = \sum_{\forall i} (p_i H_i) \sum_{\forall i} H_i^2$, [12].

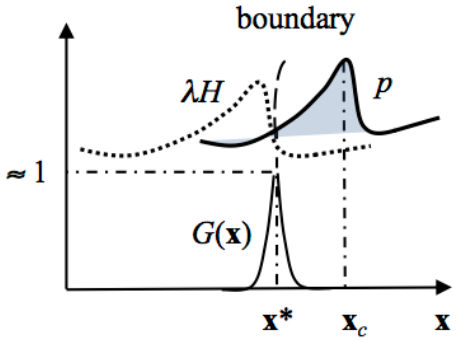

Fig. 1 *A typical plot for SBD* (p)*, SBE* (H)*, and CBI* (G)*.*

Although above definitions when combined are capable of identifying data points that reside at boundaries of clusters, they will fail in case of uniformly distributed data points and non-convex datasets. They also disregard intra-cluster data scatter behavior, which may yield incorrect cluster-like regions. Thus, it is not obvious how distinct clusters will be associated to overall data scatter rather than inter-cluster data points. In following section, we provide a new approach to remedy these issues and to identify cluster-like regions with combined scatter and data associative characteristics.

3. New clustering algorithm

Given a dataset of N data points in a D-dimensional hyperspace, a similarity measure between $\mathbf{x}_i$ and $\mathbf{x}_j$ and similarity-based probabilistic distribution for $\mathbf{x}_i$ were defined in Eq. 1 and 2, respectively. It is seen that similarity definition in Eq. 1 can be interpreted as a joint probabilistic term between $\mathbf{x}_i$ and $\mathbf{x}_j$, that is $s_{ij} = p_{i,j} = p(\mathbf{x}_i, \mathbf{x_j})$. Hence, it concludes that the SBD description in Eq. 2 refers to a marginal probability term while SBE in Eq. 5 can be interpreted as a similarity-based marginal entropy for data point $\mathbf{x}_i$.

In order to associate data points to representative cluster identifiers in forming clusters, we need to derive an association metric in terms of proximities between data points. We suggest a conditional entropy between $\mathbf{x}_i$ and $\mathbf{x}_j$ as

$$H_{i|j} = H(\mathbf{x}_i \,|\, \mathbf{x}_j) = p_{i,j} \log (p_j / p_{i,j}) = -\, s_{ij} \log s_{ij} + \, s_{ij} \log p_j \tag{6}$$

to express measure of association. With above conditional entropy defintion, it is appropriate to identify how two data points can be related to each other in terms of being representative on each other since $H_{i|j} \neq H_{j|i}$. We consider the following expression of ambiguity between them given by

$$|\Psi(\mathbf{x}_i, \mathbf{x}_j)| = \left|H_{i|j} - H_{j|i}\right| = |s_{ij} \log(p_j/p_i)|\,, \tag{7}$$

which is a measure of net uncertainty reduction due to respective neighborhood between these two data points. Since $|\Psi(\mathbf{x}_i, \mathbf{x}_j)|$ is a reflexive, i.e. symmetric, relationship, it is a metric and can be used to assess how two distinct points convey information for each other. Total net uncertainty with distinctiveness for $\mathbf{x}_i$ to be a representative data point within a grouping can be expressed as

$$H_\Delta\left(\mathbf{x}_i\right) = \left|\sum_{\forall j \neq i} \Psi(\mathbf{x}_i, \mathbf{x}_j)\right| = \left|p_i \log p_i + \sum_{\forall j \neq i} s_{ij} \log p_j\right|, \tag{8}$$

which is a non-negative quantity for any i with 0 as a global minimum. For a simplified view of description above, let us consider the case where $\mathbf{x}_i$ is within a fixed distance to other data points such that $s_{ij,j\neq i} = \delta$, i.e. $p_i = (N-1)\delta$ and $s_{jk,k\neq j} = \gamma$, i.e. $p_j = (N-1)\gamma$, then $H_\Delta \approx |-(N-1)\delta \log?(\gamma/\delta)|$. If we assume that $\mathbf{x}_i$ is within a region of data points closely localized to each other and $\mathbf{x}_i$, i.e. $\delta \approx \gamma \approx 1$, then we should expect $H_\Delta(\mathbf{x}_i)$ to decay to a local minimum. If other data points are taken further away from $\mathbf{x}_i$, then δ becomes smaller, which may then increase $H_\Delta(\mathbf{x}_i)$. However, depending on scatter properties of other neighboring data points such that $\delta > \gamma$, $H_\Delta(\mathbf{x}_i)$ will decrease. Thus, we conclude that $\mathbf{x}_i$ is a possible cluster identifier with p_i which goes to a locally maximum enforcing $H_\Delta(\mathbf{x}_i)$ to a locally minimum, e.g. 0. Hence, it is possible to suggest an indicator function such as $\theta_i = e^{-[H_\Delta(\mathbf{x}_i)]^2}$ to identify cluster-related points: θ_i may be exploited to refer to availability of a cluster when it is greater than a suitably chosen threshold θ_{th}, i.e. $\mathbf{x}_i$ being a cluster identifier if $\theta_i > \theta_{th}$. Each candidate cluster is initialized with a respective identifier $\mathbf{c}$ found as such. Cluster membership for other points can be assigned to individual identifiers by using a quantity called *mutual ambiguity* given by $\Psi(c, x_j) + H_\Delta(c)$ where $\mathbf{c} \neq \mathbf{x}_j$: Given a set of N data points $\mathbf{X}= \{\mathbf{x}_i\}_{i=1\ldots N}$ and a set of K cluster identifiers $\mathbf{C} = \{\mathbf{c}_k\}_{k=1\ldots K}$, a data point $\mathbf{x}_m \in \mathbf{X}$ is assigned to a cluster on the basis of minimum mutual ambiguity

$$\arg\min_k \left|\Psi\left(\mathbf{c}_k, \mathbf{x}_m\right) + H_\Delta\left(\mathbf{c}_k\right)\right|, \tag{9}$$

once possible cluster identifiers have been identified. It is seen that the main computation of the new algorithm is due to calculation of similarity and similarity-based sample marginal probability terms in forming Eq. 2, 5–7 along with the indicator function to determine possible cluster centers by using Eq. 9. Once these terms have been obtained, it is straightforward to cluster data points in kN steps. Since these data points can be determined a priori without further iteration or recursion, the algorithm is single-step.

Having described the new algorithm, it is convenient to demonstrate its capabilities and compare it against a popular partitioning algorithm, e.g. k-means,

for clustering a dataset with two non-convex ring-shaped inner-clusters shown in Fig. 2(a). For comparison purpose, a variant of k-means algorithm called k-means* [11], which adopts weight adjustment of clusters is chosen. The k-means* algorithm is required to be provided with the number of clusters available *a priori*. On the other hand, for the new algorithm, the cluster identifiers are found with $\theta_i > \theta_{th}$ where θ_{th} is taken 0.95. A set of 2000 vectors $\mathbf{x} = [x_1 x_2]$ was generated from uniformly distributed 2D (bivariate) random density within region $|x_{1,2}| \leq 2.5$ and then the respective circular regions were defined as clusters. As an example, Fig. 2(a) shows two such clusters while Fig. 2(b) and (c) visualize simulation results for k-means* and new algorithm, respectively.

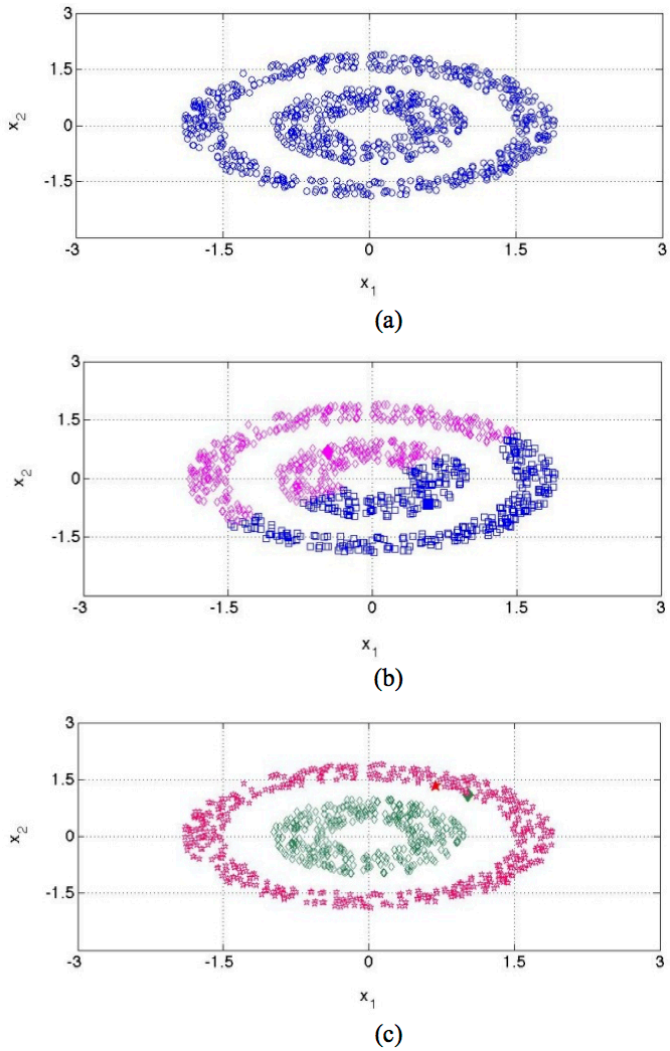

Fig. 2 *Data scatter of two inner-clusters: (a) raw dataset, (b) clustered with k-means* [11], (c) clustered with new algorithm.*

From above plot, it is seen that the new method successfully extracts the original non-convex clusters without distortion. On the other hand, k-means* counterpart fails even in regenerating shape of the clusters.

4. Experiments

Further to justification above, two sets of 100 experiments were conducted with publicly available real datasets for comparing new method and its information-theoretic hierarchical (normal-density) model-based [1], DBSCAN [5], kernel spectral [8], k-means* [11], hierarchical splitting [13] counterparts. The kernel width for spectral method and other relevant methods was taken 1. For k-means*, spectral and information-theoretic model-based algorithms, the initial number of clusters was taken twice the (actual) number of clusters (or classes) and iteratively decremented to 1 with randomly selected data points ac centroids, which is an hindering factor. The number of clusters was estimated based on Davies–Bouldin index [7], as a compactness factor. Densities for DBSCAN and the clusters for splitting methods were initialized with data points having similarity larger than 0.5 instead of conditional constraint of minimum number of samples to initiate density formation. In evaluation, a pre-specified quadratic mutual-information function for inclusion/exclusion of data points with randomly initialized clusters built. Those data points that add/subtract to incremental variation were included/excluded for respective cluster availability. The algorithms were also evaluated in statistical terms in (number of successfully classified data points)/N and (number of iterations)/N *once* number of classes/clusters has been found successfully.

For the first set of 100 experiments, the Character Trajectories Dataset [18] was used. Dataset consists of 3-dimensional 2858 labeled data points of pen tip segment trajectories for the 20 single pen-down characters, e.g. 'a', 'e', 'w'. The feature vectors are composed of respective coordinates x, y, and pen tip force. At each experiment, 50 random samples from each of randomly selected 5 characters were drawn, i.e. $N = 250$. Minimum number of samples for DBSCAN algorithm to initiate density formation was taken 25. Results of the algorithms studied with this dataset are summarized in Tab. I.

The second set of experiments was carried out with use of Musk (Version 2) Dataset [9]. This dataset describes a set of 102 molecules of which 39 are judged by human experts to be musks and the remaining 63 molecules are judged to be non-musks. However, the $D = 166$ features that describe these molecules depend upon the exact shape, or conformation, of the molecule. Because bonds can rotate, a single molecule can have different shapes. To generate this data set, all the low-energy conformations of the molecules were generated to produce 6598 conformations. At each experiment, 100 random data points were drawn from each class, i.e. $N = 200$. Each feature was normalized to respective maximum to allow straightforward computation of distances and avoid prohibitive matrix inversion operation. Simulation results of the algorithms studied with the second dataset are summarized in Tab. II.

From Tab. I and II, it is seen that for both datasets, the new clustering algorithm exhibits better and consistent statistical characteristics in (correctly) estimating the number of clusters and scattering data points to estimated clusters/classes

Algorithm	Classification success [%] Avg. / Std. dev.	Number of clusters found Avg. / Std. dev.	(Number of iterations)/N Avg. / Std. dev.
New	68.3 / 2.8	4.8 / 0.7	1.4 / < 0.2
Inform. theoretic kernel density, [1]	43.3 / 4.6	5.7 / 1.9	414.3 / 15.2
DBSCAN, [5]	57.5 / 3.2	4.2 / 1.6	73.6 / 7.5
Kernel spectral, [8]	44.2 / 4.1	4.5 / 1.5	295.3 / 11.2
k-means*, [11]	53.6 / 3.2	5.3 / 1.5	92.1 / 7.8
Splitting, [13]	49.7 / 3.5	4.4 / 1.3	302.5 / 11.7

Tab. I *Statistical performance measures for the proposed (New) and some other clustering algorithms with Character Trajectory Dataset.*

Algorithm	Classification success [%] Avg. / Std. dev.	Number of clusters found Avg. / Std. dev.	(Number of iterations)/N Avg. / Std. dev.
New	57.1 / 3.1	2.7 / 0.8	5.2 / <1.6
Inform. theoretic kernel density, [1]	42.8 / 5.1	4.9 / 1.6	421.8 / 19.6
DBSCAN, [5]	54.2 / 4.1	3.4 / 1.2	218.4 / 11.5
Kernel spectral, [8]	46.5 / 5.0	3.5 / 1.6	1989.8 / 27.6
k-means*, [11]	39.1 / 3.9	3.7 / 1.9	371.6 / 12.2
Splitting, [13]	43.6 / 4.2	3.9 / 1.5	19140 / 82.4

Tab. II *Statistical performance measures for the proposed (New) and some other clustering algorithms with Musk (Version 2) Dataset,* $D = 166$.

with much smaller number of iterations compared its counterparts. High classification success, which is the ratio of data points correctly assigned to their respective classes *once* the correct value of number of clusters/classes has been established results from inherent associative nature of the new algorithm between data points in dataset. Of other counterparts, DBSCAN exhibits closer performance of classification success for both datasets to the new algorithm than others while its time complexity is still much higher than the new algorithm. As expected, splitting algorithm has the highest time complexity for the second dataset due to increased feature vector dimension.

5. Conclusions

In this paper, a single-step clustering algorithm is described based on a new information-theoretic association measure that uses a quantity called ambiguity metric as a net conditional entropy difference between data points. The new algo-

rithm identifies possible clusters based on how density and scatter characteristics of neighboring data points around each data point vary associated to well-defined similarity description. The algorithm is fully unsupervised, i.e. no requirement of a priori knowledge of number of clusters or shape of clusters as opposed to major counterparts is imposed. Since all computation is performed in single step, no recursion or iteration is required. It is illustrated that new algorithm is capable of extracting even non-convex clusters with their actual shapes for which most major algorithms fail. Simulations with two real datasets for new algorithm and some major counterparts show that the new method statistically outperforms the counterparts in terms of successful identification of clusters and assigning data points to respective clusters/classes with much lower time complexity.

References

[1] AGHAGOLZADEH M., ZADEH A.S., ARAABI B.N. Information Theoretic Hierarchical Clustering, Entropy, 2011, 13, pp. 450–465.

[2] AMORIM R.C., MIRKIN B. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering, Pattern Recognition, 2012, 45(3), pp. 1061–1075, doi: 10.1016/j.patcog.2011.08.12.

[3] AZZALNI A., TORELLI N. Clustering via nonparametric density estimation, Statistical Computation, 2007, 17, pp. 71–80, doi: 10.1007/s11222-006-9010-y.

[4] BUHMANN J.M. Information theoretic model verification for clustering, Proc. IEEE Int. Symp. Information Theory, 2010, pp. 1398–1402.

[5] CAMPELLO R.J.G.B., MOULAVI D., SANDER J. Density-Based Clustering Based on Hierarchical Density Estimates, Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, 2013, 7819, pp 160–172, doi: 10.1007/978-3-642-37456-2_14.

[6] CASTRO E.A., CHEN G., LERMAN G. Spectral clustering based on local linear approximations, Elect. J. of Statistics, 2011, 5, pp. 1537–1587, doi: 10.1214/11-ejs651.

[7] DAVIES D.L., BOULDIN D.W. A Cluster Separation Measure, IEEE Trans. Pattern Analysis and Machine Intelligence, 1979, PAMI-1(2), pp. 224–227, doi: 10.1109/TPAMI.1979.4766909.

[8] DHILLON I.S., GUAN Y., KULIS B. Kernel k-means, Spectral Clustering and Normalized Cuts, Proc. Int. Conf. Konwledge Discovery and Data Mining, (ACM SIGKDD), 2004, pp. 551–556.

[9] DIETTERICH T. UCI Machine Learning Repository, Irvine, CA, Musk (Version 2) Data Set, https://archive.ics.uci.edu/ml/datasets/Musk+{%}28Version+2{%}29, 1994, AI Group at Arris Pharmaceutical Corporation, CA, USA.

[10] JENSSEN R., HILD K.E., ERDOGMUS D., PRINCIPE J.C., ELTOFT T. Clustering using Renyis entropy, Neural Networks, 2003, 1, pp. 523–528.

[11] MALINEN M.I., ISTODOR R.M., FRÄNTI P. K-means*: Clustering by gradual data transformation, Pattern Recognition, 2014, 47(10), pp. 3376–3386, doi: 10.1016/j.patcog.2014.03.034.

[12] RASMUSSEN C.E. The Infinite Gaussian Mixture Model, Advances in Neural Information Processing, S. A. Solla, T. K. Leen, and K.–R. Muller, Eds. MIT Press, 2000, pp. 554–560.

[13] SLINK S.R. An optimally efficient algorithm for the single-link cluster method, The Computer J., 1973, 16(1), pp. 30–34, doi: 10.1093/comjnl/16.1.30.

[14] SUGIYAMA M., NIU G., YAMADA M., KIMURA M., HACHIYA H. Information-maximization clustering based on squared-loss mutual information, Neural Computation, 2014, 26(1), pp. 84–131. doi: 10.1162/NECO_a_00534.

[15] TEMEL T. Finding number of clusters in single-step with similarity-based information-theoretic algorithm, IET Elect. Lett., 2014, 50(1), pp. 29–30, doi: 10.1049/el.2013.3362.

[16] TEMEL T., AYDIN N. A threshold free clustering algorithm for robust unsupervised classification, Proc. ECSIS Symposium on Bio-inspired, Learning, and Intelligent Systems for Security, BLISS, (IEEE), 2007, pp. 119–122, doi: 10.1109/BLISS.2007.26.

[17] TUZEL O. Kernel methods for weakly supervised mean shift clustering, Proc. IEEE Int. Conf. Computer Vision, 2009, pp. 48–55, doi: 10.1109/ICCV.2009.5459204.

[18] WILLIAMS B.L. UCI Machine Learning Repository, Irvine, CA, Character Trajectories Data Set, http://archive.ics.uci.edu/ml/machine-learning-databases/character-trajectories/, 2008, School of Informatics, University of Edinburgh, UK.