ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

A New Algorithm for Clustering in Single-step based on an Information-theoretic Mutual Irrelevance Metric

Turgay TEMEL (PhD)
Department of Mechatronics Engineering, Bursa Technical University, Bursa, TURKEY
turgay.temel@btu.edu.tr / ttemel70@gmail.com

Manuscript Info

Manuscript History:

Received: 14 May 2015
Final Accepted: 22 June 2015
Published Online: July 2015

Key words:

Clustering; information theory; cluster analysis; machine learning; unsupervised learning

Corresponding Author

Turgay TEMEL (PhD)

Abstract

A new algorithm that can extract clusters in single-step based on a new information-theoretic notion is described. New method employs similarity-based sample entropy and probability descriptions to express scatter in a given dataset. Based on these quantities, a new information-theoretic association measure called mutual irrelevance metric is defined to model a (dis)-connectivity rule between samples. This metric is utilized for determining candidate cluster representative samples coined cluster indicators. Possible clusters are established based on an association quantity between samples and cluster indicators in a single iteration. Clustering capability of new approach is demonstrated for a non-convex dataset, which is hard to cluster by using most well known counterparts. It is also tested and compared to major algorithms for publicly available real datasets. Experimental results reveal that the proposed approach outperforms predecessors it is compared to.

Introduction

Given a dataset, a clustering algorithm is judged in terms of its capability in identifying most representative clusters as a means of compact groupings with optimal fit to sample scatter, which can be validated in varying ways, [1]. Prominent clustering methods seek certain regularities in scatter of possible cluster samples which can be attributed to by statistical models: Partitioning methods search for an optimal partition with k clusters to which N data points in dataset in feature space are to be assigned based on a criterion as a separation quality between clusters against compactness within clusters. They are of iterative nature to obtain such a factor of variation in distortion/compactness measure versus the *a priori* number for clusters. For example, k-means, [2], and mixture of varying densities/distributions, [3], generally run with complexity of O(kNI) where k is the *a priori* number of clusters assumed and I is the number of iterations which heavily relies upon a prespecified termination rule. As a different group of partitioning algorithms, spectral methods can successfully extract non-convex clusters in dataset with reduced dimensions by using k largest eigenvectors corresponding to Laplacian of the similarity or adjacency matrix, [4]. They operate with complexity of O($N^{3/2}$+kNI) or higher where initialisation and/or post processing steps may benefit from some fast partitioning methods, such as k-means. Hierarchical methods form clusters by merging/dividing sample subgroupings recursively till no more connected group is possible, which yields a connectivity pattern called dendogram. Due to intensive recursions, their time complexity is high, e.g. O(N^3) while time complexity of some speed-up agglomerative hierarchical algorithms reduces to O(N^2), [5]-[6]. Although they are generally not good competent at identifying overlapping densities as well as sensitive to outliers, most hierarchical algorithms do not require the *a priori* knowledge for the number of clusters. Density-based clustering algorithms determine clusters as sample groups where they are densely localized. DBSCAN, [7], similar to

connectivity in hierarchical clustering algorithms, assumes clusters from those connected samples which reside at a distance smaller than a pre-specified threshold as well as subject to a minimum number of samples which satisfy the former criterion. In mean-shift clustering, individual samples are (virtually) relocated or replicated to the possibly densest neighbourhood based on estimated maxima of kernel density, [8]. A sample that resides at a (local) maximum of density after samples have moved is considered as a centroid/mean. Both density-based methods do not involve the knowledge for the number of clusters *a priori*, and impose no constraint on the shape of the clusters. Since mean-shift algorithm highly relies on estimating the neighboring samples to which mean vectors are to be shifted at each successive step, it is computationally expensive with complexity of $O(N^2I)$ where I was cited previously compared to DBSCAN which usually has complexity of $O(N\log N)$.

Information-theoretic entropy and mutual-information descriptions are invariant to modelling data representation besides capturing higher-order statistics involved, [9]. Since they also model sample scatter properties well where entropy and mutual information are considered as intuitive association measures, these have been employed in clustering, e.g. kernel-based hierarchical clustering with use of optimized quadratic mutual-information in [10], and clustering algorithm based on Renyi's entropy in [11]. A recent study in [12] describes a method to estimate the number of clusters in single-step by identifying samples at possible cluster boundaries in a dataset based on information-theoretic sample entropy and probability descriptions. However, to our knowledge, there is no algorithm that gives both the number of clusters and forms the respective groupings in single-step or one-pass.

This study presents a new, single-step information-theoretic algorithm based on a quantity called irrelevance metric between samples with use of similarity-based sample entropy and probability descriptions. Simulations show that new algorithm is highly successful in clustering even non-convex datasets for which major clustering algorithm fail in extracting actual cluster shapes. Furthermore, it is tested and compared to major algorithms for publicly available real datasets. Experimental results reveal that the proposed approach outperforms predecessors it is compared to in statistical terms.

Proximity-based Sample Density and Entropy Definitions

A definition of similarity, s_{ij}, which is also used in this study, between D-dimensional data samples (or feature vectors) $\mathbf{x}_i$ and $\mathbf{x}_j$ is commonly defined as

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

The distance metric is Euclidean-squared distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ given by $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$. The parameter β is the kernel size or resolution parameter and it is usually taken 1. Then, given a dataset of N samples, we can define a similarity-based (experimental) probabilistic (SBP) term, to which $\mathbf{x}_i$ is at the center of respective kernel density, as

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

The parameter γ_i is chosen such that, $\sum_{\forall j \neq i} \gamma_i = 1$, which will be disregarded in this study. 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 vector $\mathbf{x}_i$ - $\mathbf{x}_j$ and '•' is the dot-product operator. It should be noted that valley-seeking and mean-shift clustering algorithms seek a sample point $\mathbf{x}_i$ that meets the following condition

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

It is seen that as other samples are brought closer to $\mathbf{x}_i$ respective SBD reaches a local maximum. In this manner, SBD is not a suitable quantity to assess sample scatter properties merely in uncertain or uniformly distributed

regions, which can be considered as boundaries between clusters. For example, in overlapping regions, due to superposition of similarities, SBD will not be able to adequately represent scatter of samples around $\mathbf{x}_i$. Similar to stochastic entropy definition, an experimental counterpart can be devised to express data scatter for this purpose, which we call similarity-based sample entropy (SBE). Thereby, it would be desirable to situate regions where samples are difficult to infer for membership to candidate clusters. It is possible to introduce a similarity-based entropy for sample $\mathbf{x}_i$ as

$$H_i = -\sum_{\forall j \neq i} s_{ij} \log s_{ij} \tag{5}$$

From above descriptions, it is straightforward to observe close resemblance between information-theoretic probabilistic, i.e. stochastic, entropy and its experimental similarity-based counterpart: Given a particular sample $\mathbf{x}_i$ its SBE decreases to a minimum as other samples are moved either very close to or far from it. It reaches a maximum as they are brought at a distance irregularly or randomly, which makes it difficult to assert on closeness.

Proposed Clustering Algorithm

For a dataset of N samples (or feature vectors), $\mathbf{x}_i$ and $\mathbf{x}_j$ in a D-dimensional data hyperspace, the similarity-based probability (SBP) for $\mathbf{x}_i$ defined in (2) can be regarded as a marginal probability term where $s_{ij} = p_{ij} = p(\mathbf{x}_i, \mathbf{x_j})$ is the joint probability between $\mathbf{x}_i$ and $\mathbf{x}_j$. Similarly, having defined the sample SBP as per, it is convenient to introduce similarity-based sample marginal entropy (SBE) as

$$H_i = H(\mathbf{x}_i) = -\sum_{\forall j \neq i} s_{ij} \log s_{ij} \tag{6}$$

However, as suggested by [12], although SBP and SBE are capable of identifying samples that reside at boundaries of clusters in single-step, they fail in representing association to possible groupings. In order to associate samples to representative cluster identifiers, for a possible grouping or clustering, an association metric in terms of proximities between samples needs to be derived. Then, the conditional entropy between samples $\mathbf{x}_i$ and $\mathbf{x}_j$ can be given by

$$H_{i|j} = H(\mathbf{x}_i \mid \mathbf{x}_j) = p_{ij} \log(p_j / p_{ij}) = -s_{ij} \log s_{ij} + s_{ij} \log p_j \tag{7}$$

With conditional entropy definition in (7), it is possible to express how two samples can be interrelated in terms of a metric since $H_{i|j} \neq H_{j|i}$. For this purpose we define a quantity called mutual irrelevance metric as a difference

$$\left|\Psi(\mathbf{x}_i, \mathbf{x}_j)\right| = \left|H_{i|j} - H_{j|i}\right| = \left|s_{ij} \log(p_j / p_i)\right| \tag{8}$$

It should be noted that $\left|\Psi(\mathbf{x}_i, \mathbf{x}_j)\right|$ in (8) refers to a measure of net uncertainty reduction due to neighbourhood between these two samples. Total net uncertainty as distinctive characteristics for sample $\mathbf{x}_i$ to be a representative sample within a grouping can be expressed as

$$I_\Delta(\mathbf{x}_i) = \left|\sum_{\forall j \neq i} \Psi(\mathbf{x}_i, \mathbf{x}_j)\right| = \left|-p_i \log p_i + \sum_{\forall j \neq i} s_{ij} \log p_j\right| \tag{9}$$

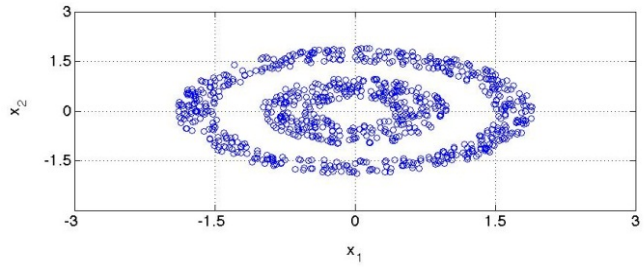
which is a non-negative quantity for any i with 0 as a global minimum. For a simplified explanation of new descriptions above, we assume that $\mathbf{x}_i$ is within a fixed distance to other samples such that $s_{ij} = \delta$, with $j \neq i$, i.e. $p_i = (N-1)\delta$ and $s_{jk} = \gamma$, for $k \neq j$, i.e. $p_j = (N-1)\gamma$, then $I_\Delta \approx |-(N-1)\delta \log(\gamma/\delta)|$. If we regard $\mathbf{x}_i$ being within a region of samples closely localized with respect to each other and $\mathbf{x}_i$, i.e. $\delta \approx \gamma \approx 1$, then we should expect $I_\Delta(\mathbf{x}_i)$ to decay to a local minimum. If other samples are moved further away from $\mathbf{x}_i$, then δ gets smaller, which may result in increase in $I_\Delta(\mathbf{x}_i)$. However, depending on scatter properties of other samples such that $\delta > \gamma$, $I_\Delta(\mathbf{x}_i)$ may decrease. Thus, we conclude that $\mathbf{x}_i$ is a possible cluster indicator with p_i, which goes to a locally maximum and enforces $I_\Delta(\mathbf{x}_i)$ to a locally minimum, e.g. 0. Similar to the definition of cluster boundary indicator function proposed in [12], an indicator function such as $\varphi(i) = e^{-[I_\Delta(\mathbf{x}_i)]^2}$ can be exploited to identify such representative samples. It is observed that $\varphi(i)$ may be exploited to refer to availability of a cluster when it is greater than a suitably chosen threshold φ_{th},

i.e. $\mathbf{x}_i$ is a cluster indicator $\mathbf{c}$ if $\varphi(i) > \varphi_{\text{th}}$. Each candidate cluster is initialised with respective indicator found as per. Other samples can be assigned to corresponding clusters by using a similar approach to mutual-information. Given a set of N samples $\mathbf{X}=\{\mathbf{x}_i\}_{i=1\ldots N}$ and a set of K cluster indicators $\mathbf{C}=\{\mathbf{c}_k\}_{k=1\ldots K}$, a sample $\mathbf{x}_m \in \mathbf{X}$ can be assigned to clusters on the basis

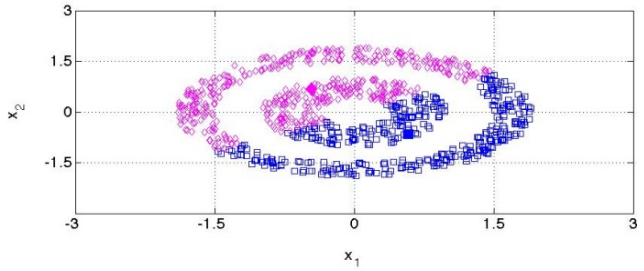
$$\underset{k}{\operatorname{argmin}} \left| \Psi(\mathbf{c}_k, \mathbf{x}_m) + I_\Delta(\mathbf{c}_k) \right| \tag{10}$$

The computational complexity of new algorithm is mainly due to calculation of similarity and similarity-based sample marginal probability terms in forming (1), (6) - (8), which easily facilitates forming the indicator function to determine possible cluster indicators. Once these terms have been obtained, (10) will provide a straightforward rule to cluster samples in kN steps once the indicator function has been formed to identify indicators. However, since these samples can be identified *a priori* without further iteration, the algorithm can be considered single-step.

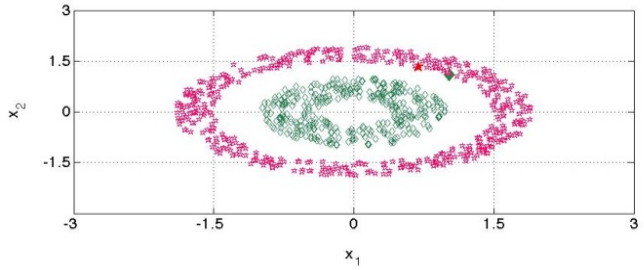
It is appropriate to visualize the capabilities of new algorithm and compare it against a popular partitioning algorithm, e.g. *k*-means. For this purpose, we consider the problem of clustering a dataset with two non-convex ring-shaped inner-clusters shown in Fig. 1(a). A variant of *k*-means algorithm called *k*-means*, [13], which adopts weight adjustment of clusters is chosen. The *k*-means* algorithm is required to be fed with the number of clusters available. For this purpose, the number of cluster indicators found by the new algorithm based on $\varphi(i) > \varphi_{\text{th}}$ with $\varphi_{\text{th}} = 0.95$ was used. 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 illustrative example, Fig. 1(a) shows two such clusters while Fig. 1(b) and (c) visualize simulation results for *k*-means* and new algorithms, respectively. From the plots, it is seen that new method successfully extracts the original non-convex clusters without distortion while *k*-means* counterpart fails even in regenerating shape of the clusters.



(a)



(b)



(c)

Fig. 1. Sample scatter of two inner-clusters: (a) raw dataset, (b) clustered with *k*-means* [13], (c) clustered with new algorithm.

Experiments

Further to justification above, two sets of 100 experiments were conducted with publicly available real datasets for comparing new method and its hierarchical splitting, [5], *k*-means*, [13], kernel spectral with *k*-means, [14], DBSCAN, [7], and information-theoretic hierarchical (normal-density) model-based [10] 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 training samples. The number of clusters was estimated based on Davies–Bouldin index, [15], as a compactness factor. Densities for DBSCAN and the clusters for splitting methods were formed with samples having similarity larger than 0.5 instead of conditional constraint of minimum number of samples to initiate density formation. For the information-theoretic algorithm, kernels utilized unity width. In evaluation, a pre-specified quadratic mutual-information function for inclusion and exclusion of samples and randomly initialized clusters were constructed. Those samples that contributed to incremental variation were included for the respective cluster otherwise excluded for considering to other available clusters. Clustering performance of algorithms were evaluated in statistical measures in (number of successfully classified samples)/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 at (http://archive.ics.uci.edu/ml/machine-learning-databases/character-trajectories/) was used. Dataset consists of 3-dimensional 2858 labelled samples 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. Performances of the algorithms studied are summarized in Table I.

Algorithm	Classification success, %	Number of clusters found	(Number of iterations)/N
	Avg. / Std. dev.	Avg. / Std. dev.	Avg. / Std. dev.
New	68.3 / 2.8	4.8 / 0.7	1.4 / < 0.2
Splitting, [5]	49.7 / 3.5	4.4 / 1.3	302.5 / 11.7
k-means*, [13]	53.6 / 3.2	5.3 / 1.5	92.1 / 7.8
Kernel spectral, [14]	44.2 / 4.1	4.5 / 1.5	295.3 / 11.2
DBSCAN, [7]	57.5 / 3.2	4.2 / 1.6	73.6 / 7.5
Information theoretic kernel density, [10]	43.3 / 4.6	5.7 / 1.9	414.3 / 15.2

Table I. Statistical performance measures for the proposed (New) and some other clustering algorithms with Character Trajectory Dataset, $D = 3$.

The second set of experiments was carried out with use of Musk (Version 2) Dataset at https://archive.ics.uci.edu/ml/datasets/Musk+%28Version+2%29. 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 samples 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. Table II summarizes performances of the algorithms studied.

Algorithm	Classification success, %	Number of clusters found	(Number of iterations)/*N*
	Avg. / Std. dev.	Avg. / Std. dev.	Avg. / Std. dev.
New	57.1 / 3.1	2.7 / 0.8	5.2 / < 1.6
Splitting, [5]	43.6 / 4.2	3.9 / 1.5	19140 / 82.4
k-means*, [13]	39.1 / 3.9	3.7 / 1.9	371.6 / 12.2
Kernel spectral, [14]	46.5 / 5.0	3.5 / 1.6	1989.8 / 27.6
DBSCAN, [7]	54.2 / 4.1	3.4 / 1.2	218.4 / 11.5
Information theoretic kernel density, [10]	42.8 / 5.1	4.9 / 1.6	421.8 / 19.6

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

Conclusions

A new information-theoretic algorithm that can identify clusters in single-step is introduced based on similarity-based sample entropy and probability descriptions. Based on these quantities, a new information-theoretic notion called mutual irrelevance metric is defined to model mutual neighborhood association between samples. This metric is further employed in identifying candidate cluster representative samples coined cluster indicators. Candidate clusters are formed based on an association quantity between samples and cluster indicators in a single iteration. New approach is justified for a non-convex dataset, which is hard to cluster by using most counterparts. Proposed and some major algorithms are tested and compared to major algorithms for publicly available real datasets. Experimental results show that the new algorithm excels the predecessors it is compared to.

References

[1] J. M. Buhmann, “Information theoretic model verification for clustering,” Proc. IEEE Int. Symp. Information Theory, pp. 1398-1402, 2010.

[2] R. C. Amorim, and B. Mirkin, “Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering,” Pattern Recognition, vol. 45, no. 3, pp. 1061-1075, 2012. doi: 10.1016/j.patcog.2011.08.12

[3] C. E. Rasmussen, “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.

[4] E. A. Castro, G. Chen, and G. Lerman, “Spectral clustering based on local linear approximations,” Elect. J. of Statistics,vol. 5, pp. 1537-1587, 2011. doi:10.1214/11-ejs651

[5] R. Sibson, “SLINK: an optimally efficient algorithm for the single-link cluster method,” The Computer J., vol. 16, no. 1, pp. 30–34, 1973.doi:10.1093/comjnl/16.1.30

[6] A. Azzalni, and N. Torelli, “Clustering via nonparametric density estimation,” Statistical Computation, vol. 17, pp. 71-80, 2007. doi: 10.1007/s11222-006-9010-y

[7] R. J. G. B. Campello, D. Moulavi, and J. Sander, “Density-Based Clustering Based on Hierarchical Density Estimates,” Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, vol. 7819, pp 160-172, 2013. doi: 10.1007/978-3-642-37456-2_14

[8] O. Tuzel, “Kernel methods for weakly supervised mean shift clustering,” Proc. IEEE Int. Conf. Computer Vision, pp. 48-55, 2009. doi: 10.1109/ICCV.2009.5459204

[9] M. Sugiyama, G. Niu, M. Yamada, M. Kimura, and H. Hachiya, “Information-maximization clustering based on squared-loss mutual information,” Neural Computation, vol. 26, no. 1, pp. 84-131, 2014.doi:10.1162/NECO_a_00534

[10] M. Aghagolzadeh, A. S. Zadeh, and B. N. Araabi, “Information Theoretic Hierarchical Clustering,” Entropy, vol. 13, pp. 450-465, 2011.

[11] R. Jenssen, K. E. Hild, D. Erdogmus, J. C.Principe, and T. Eltoft, “Clustering using Renyi’s entropy,” Neural Networks, vol. 1, pp. 523-528, 2003.

[12] T. Temel, “Finding number of clusters in single-step with similarity-based information-theoretic algorithm,” IET Electronics Letters, vol. 50, no. 1, pp. 29-30, 2014.

[13] M. I. Malinen, R. M. Istodor, and P. Fränti, “K-means*: Clustering by gradual data transformation,” Pattern Recognition, vol. 47, no. 10, pp. 3376-3386, 2014. doi:10.1016/j.patcog.2014.03.034

[14] I. S. Dhillon, Y. Guan, and B. Kulis, “Kernel k-means, Spectral Clustering and Normalized Cuts," Proc. Int. Conf. Konwledge Discovery and Data Mining, (ACM SIGKDD), pp. 551-556, 2004.

[15] D. L. Davies, and D. W. Bouldin, “A Cluster Separation Measure,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, 1979. doi:10.1109/TPAMI.1979.4766909