ISSN(Online) : 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

Clustering Algorithm for Resemblance Measure through Convex Coding

Venkata Sai Sriharsha Sammeta[1], Navya Yenuganti[2], K Ram Mohan Rao[3]

Student, Computer Science, Vasavi College of Engineering, Osmania University, Hyderabad, India[1]

Student, Computer Science, Vasavi College of Engineering, Osmania University, Hyderabad, India[2]

Ph.D. in Computer Science, Professor, Vasavi College of Engineering, Osmania University, Hyderabad, India[3]

ABSTRACT:Relational information looks frequently in many machine learning applications. Relational information consists of the dyad sapient cognations (kindred attributes or dissimilarities) between each pair of implicit objects, and are customarily stored in cognation matrices and typically no other cognizance is available. Albeit relational clustering can be formulated as graph division in approximately diligences, this formulation is not adequate for general relational information. In this paper, we aim a universal model for relative clustering predicated on symmetric convex coding. The sample is applicable to all types of relational information and cumulates the subsisting graph partitioning formulation. Under this model, we derive two option bounce optimization algorithms to figure out the symmetrical convex coding fewer than two popular distance functions, Euclidian length plus extrapolated I-divergence. Experimental evaluation and theoretical analysis show the efficacy and great potential of the proposed model and algorithms.

KEYWORDS:Machine Learning, Deep Learning, Clustering, Neural Networks, Convex Coding, HMM, Log-Linear Models, Regularization, cluster validity, clustering algorithm, line symmetry, bunch analytic thinking, resemblance measure, and unattended learning.

I. INTRODUCTION

Two types of information are utilized in unsupervised discovering, characteristic plus relative information. Feature information are in the form of feature vectors and relational information consist of the pair wise cognations (homogeneous attributes or dissimilarities) between each pair of objectives, plus are conventionally stacked away in cognation matrices and typically no other cognizance is available. Albeit feature information are the most prevalent type of information, relational information have become more and more popular in many machine learning applications, such as web mining, convivial network analysis, bioinformatics, VLSI aim, plus job scheduling. Furthermore, the relational information are more general in the sense all the feature information can be transformed into relational information under a certain distance function.

The most popular way to cluster homogeneous attribute-predicated relational information is to formulate it as the graph partitioning quandary, which has been analysed for tens. Graph breakdown seeks to cut a given graph into disjoint sub graphs which represent to disjoint bunches predicated on a certain edge cut objective. Recently, graph partition with a bound cut accusative has been demonstrated to be mathematically equipment to an opportune weighted kernel k-denotes objective function (Dhillon et al., 2004; Dhillon et al., 2005).The posit abaft the graph partitioning formulation is that since the nodes within a cluster are kindred to each other, they compose a dense sub graph. However, in general this is erroneous for relational information, i.e., the clusters in relational information are not indispensably dense clusters consisting of vigorously-cognate objects.

Figure 1 shows the relational information of four clusters, which are of two variants. In diagram 1, C1 = {v1, v2, v3, v4} plus C2 = {v5, v6, v7, v8} are 2 traditional dense clusters within which objects are vigorously cognate to for each one other. still, C3 = {v9, v10, v11, v12} and C4 = {v13, v14, v15, v16} additionally form two thin bunches, among which the objects are not cognate to one other, but they are still ”identical” to one other in the sense that they are cognate to the same set of other nodes. In Web mining, this type of cluster could be a group of music ”fans” Web pages which quota the same taste on the music and are linked to the same set of music Web varlets but are not linked to for each one other (Kumar et al., 1999). Due to the paramount of identifying this type of bunches (communities), it has

ISSN(Online) : 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

been numbered as one of the five algorithmic tests in Web search engines (Hen zinger et al., 2003). Note that ye cluster construction of ye consanguinity information in Figure 1 cannot be correctly identified by graph partitioning approaches, since they probe for only dense clusters of vigorously cognate objects by cutting the given graph into sub graphs; similarly, the pristine bi-partite graph models cannot correctly identify this type of cluster structures. Note that re-defining the cognations between the objects does not solve the quandary in this situation, since there subsist both dense and sparse clusters.

If the relational information are dissimilarity-predicated, to apply graph partitioning approaches to them, we require extra efforts on opportunely transforming them into homogeneous attribute-predicated information and ascertaining that the transformation does not transmute the cluster structures in the information. Hence, it is acceptable for an algorithm to be able to name the bunch constructions no matter which type of relational information is given. This is even more than suitable in ye position whereas the backdrop erudition about the construal of the cognations is not available, i.e., we are given only a cognation matrix and do not ken if the cognations are kindred attributes or dissimilarities.

In this paper, we suggest a universal example for relative clustering predicated on symmetric convex coding of the cognation matrix. The recommended model is applicable to the general relational information consisting of only pairwise cognations typically without other cognizance; it is capable of learning both dense and sparse clusters at the same time; it amalgamates the subsisting graph partition models to provide a generalized theoretical substructure for relational clustering. Under this example, we deduce reiterative bound optimization algorithms to solve the symmetric convex coding for two consequential length occasions, Euclidian outdistance plus extrapolated I divergence. The algorithms are applicable to general relational information and at the same time they can be facilely habituated to learn a categorical type of cluster structure. For example, when enforced to discovering exclusively dumb bunches; they provide incipient efficient algorithms for graph partitioning. The convergence of the algorithms is theoretically ensured. Experimental evaluation and theoretical analysis show the efficacy and great potential of the proposed model and algorithms.

II. RELATED WORK

Graph partitioning (or clustering) is a popular expression of relative bunch, which divides the nodes of a graph into clusters by obtaning the best edge cuts of ye graph. Various sharpness cut aims, this as the average cut (Chanet al., 1993), average sodality (Shi & Malik, 2000), renormalized cut (Shi & Malik, 2000), and min-max cut (Dinget al., 2001), have been proposed. Sundry apparitional algorithms have got constituted formulated for these Documentary operates (Chan e t al., 1993; Shi & Malik, 2000; Ding et al., 2001). These algorithms utilize the eigenvectors of a graph chemical attraction matrix, or a matrix derived from the affinity matrix, to partition the graph.

Multilevel methods have been expended extensively for graph breakdown with ye Kernighan-Lin objective, which endeavor to minimize the cut in ye graph while keeping equal-sized bunches.

Recently, graph partitioning with an edge cut objective has been shown to be mathematically equipollent to a felicitous weighted kernel k-betokens objective function (Dhillon et al., 2004; Dhillon et al., 2005). Predicated on this parity, the weighted kernel k-betokens algorithm has been suggested for graph segmentation (Dhillon et al., 2004; Dhillon et al., 2005). Yu et al. (2005) propose the graph factoring bunch for ye graph partition, which attempts to conception a bipartite graph to approximate an afforded graph.

In this paper, our focus is on the homogeneous relational information, i.e., the objects in the information are of the same type. There are approximately attempts in ye lit that can be considered as clustering heterogeneous relational information, i.e., variants of objects are cognate to each other. For example, co clustering addresses clustering two types of cognate objects, such as documents and words, concurrently. Dhillon et al. (2003) recommended a co-clustering algorithm to maximize ye common information. A more generalized co-clustering framework is presented by Banerjee et al. (2004) wherein any Bregman divergence can be utilized in the objective function. Long et al., Li and Ding et al. all model the co-clustering as an optimization quandary involving a triple matrix factorization.

ISSN(Online) : 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

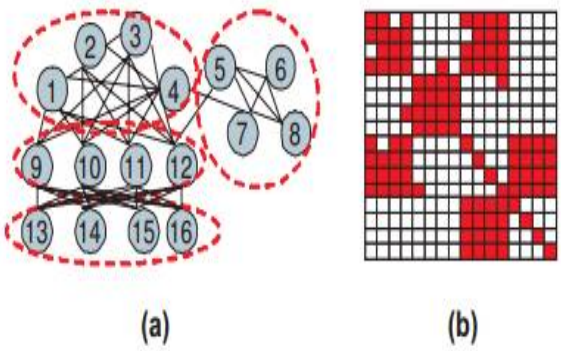

Figure 1. The graph (a) and relation matrix (b) of the relational information with different types of clusters. In (b), ye dark color announces 1 plus ye dismount color denote 0.

III. IMPLEMENTATION

A. *THE LINE SYMMETRY DISTANCE AND THE VALIDITY MEASURE USING LINE SYMMETRY:*

What is line symmetry? For a 2-dimenstional figure, if it can be folded in such a way that one-a moiety of it lies precisely on the other half is verbalized to have line symmetry. The conception of line symmetry is very pellucid and simple but an immediate quandary is how to find a metric to quantify line symmetry. A kind of line symmetry distance was proposed in. In this approach, the symmetrical line of a information set is defined by a center vector and an angle between the major axis of the information set and the x axis. The information of the major axis of the information points belonging to a class or a cluster is computed by the moment of order (p + q) method. Then the major axis is treated as the symmetrical line of that class or cluster. Furthermore, the Kd-tree predicated most proximate neighbour search is utilized to reduce the intricacy of computing the line distance. It is not pellucid how the proposed line symmetry distance can be generalized to high-dimensional space. In their approach, the symmetrical line is defined by a point and an angle between the major axis and the x axis. As we ken, a point amalgamated with an angle can define one and only one line in 2-dimensional space but will engender an illimitable number of lines in high-dimensional space. In advisement, the authors themselves verbally expressed that the performance of their line symmetry distance depends on the cull of the number of most proximate neighbours' utilized in the computation of line symmetry distance. The authors in proposed another homogeneous definition of line symmetry in. They surmised that a information set is symmetrical then it should be withal be symmetric with deference to the first principal axis of the information set. This postulation may not be hold for many situations. There are objects which is symmetrical with respective to their second principal axis but not symmetrical to their first principal axis. Consequently, an incipient line symmetry distance which can deal with high-dimensional information points is developed by our precedent work [46].

B. *THE LINE SYMMETRY DISTANCE:*

In one of our antecedent bring, a soi-distant "point symmetry" distance (symmetry about a cluster center) was proposed in. In, Bandyopadhyay and Saha proposed an incipient point symmetry-predicated distance measure. Their algorithm applies GA and Kd-tree predicated most proximate neighbor search to reduce the involution of finding the most proximate symmetric point. Albeit object with point symmetry is very widespread, line symmetry is the most mundane type of symmetry around us. Predicated on the point symmetry distance, an incipient line symmetry distance is developed in our work. Afore we present the definition of the proposed line symmetry distance, we briefly review the definition of the point symmetry distance.

A 2-dimensional figure is with point symmetry if it can be rotated 180 degrees about a point onto itself. To generalize this conception of point symmetry to quantify the degree of point symmetry for a set of high-dimensional information patterns, a point symmetry distance is defined as follows. Given a information set X containing N patterns, xj, j = 1… N, and an address transmitter c (e.g. a cluster center), the "point correspondence outdistance" between a pattern xj and the reference vector c is defined as

ISSN(Online) : 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

$$d_{ps}(\underline{x}_j, \underline{c}) = \min_{\substack{i=1,\cdots,N \\ and\ i \neq j}} \frac{\| (\underline{x}_j - \underline{c}) + (\underline{x}_i - \underline{c}) \|}{(\| \underline{x}_j - \underline{c} \| + \| \underline{x}_i - \underline{c} \|)}. \quad (4)$$

Note that Eq. (4) is minimized (i.e., dps (xj, c) = 0) if there is a pattern x ps xj) □j* = (2c subsists in the information set (visually perceive Fig. 2). The pattern x ps j* is then denoted as the point symmetrical information pattern relative to xj with veneration to c as shown in Fig. 2.

IV. EXPERIMENTAL RESULTS

We illustrate the efficacy of the proposed validity assess by trying some information sets with unlike geometrical structures. For the comparison purport, these information the partition coefficient sets were additionally tested by the three popular validity measures (PC), the relegation entropy (CE) and the Xie-Beni's disunion measure (S). The FCM algorithm or the Gustafson-Kessel (GK) algorithm [7] is employed to bunch these information sets at for each one cluster number c from c = 2 to c = 10. According the experimental results, we descry that the FCM algorithm is not felicitous to be applied for ellipsoidal clusters, but the GK algorithm can be habituated to cluster spherical and ellipsoidal clusters. Consequently, in examples 2 to 5, the GK algorithm is applied to cluster the information patterns. A value of the fuzzy exponent m = 2 was culled for the FCM algorithm and the GK algorithm. The parameter d0 was opted to be 0.005 for the composite line symmetry distance.

Example 1: The information set shown in Fig. 1 (a) consists of an amalgamation of a spherical cluster and two linear clusters. We engendered the information patterns desultorily with uniform distribution. The total number of information patterns is 400. In this example, the FCM algorithm is applied to cluster the information patterns. The total number of information patterns is 400. The performance of each validity measure is tabulated in Table 1. In Table 1 and others to follow, the highlighted (bold and shaded) ingressions correspond to optimal values of the quantifications. Note that only the LS validity measure finds the optimal cluster is three, but the PC, CE and S validity measures optate two clusters as the optimal partition. The bunch result accomplished by the FCM algorithm at c = 3 is shown in Fig. 1 (c).

Table 1. Numerical values of the validity measures for Example 1.

C	2	3	4	5	6	7	8	9	10
PC	**0.923**	0.877	0.790	0.829	0.792	0.723	0.738	0.674	0.661
CE	**0.153**	0.242	0.397	0.358	0.440	0.546	0.573	0.666	0.709
S	**0.048**	0.054	1.372	0.075	0.142	0.297	0.363	0.257	0.218
LS	0.050	**0.017**	0.037	0.096	0.238	0.170	0.160	0.184	0.158

Example 2: We desultorily engendered a heap of spherical and ellipsoidal clusters with Gaussian distribution. This information set consists of 850 information patterns distributed on five clusters, as shown in Fig. 5 (a). The clustering results achieved by the GK and FCM algorithms at c = 5 are shown in Figs. 5 (b) and (c). We descry that the FCM algorithm is not opportune to be applied for ellipsoidal clusters; the two non-overlapped compact clusters are not clustered correctly. Ergo, in this example, the GK algorithm is applied to cluster the information patterns. The performance of each validity measure is given in Table 2. Both the S and LS validity measures find that the optimal cluster number c is at c = 5, but both the PC and CE validity measures denote two clusters as the optimal partition.

Table 2. Numerical values of the validity measures for Example 2.

C	2	3	4	5	6	7	8	9	10
PC	**0.836**	0.743	0.738	0.780	0.731	0.681	0.654	0.634	0.591
CE	**0.287**	0.469	0.524	0.484	0.592	0.698	0.771	0.840	0.911
S	0.398	0.180	0.186	**0.082**	0.383	0.392	0.296	0.274	0.796
LS	0.143	0.083	0.045	**0.041**	0.091	0.152	0.126	0.180	0.137

ISSN(Online) : 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

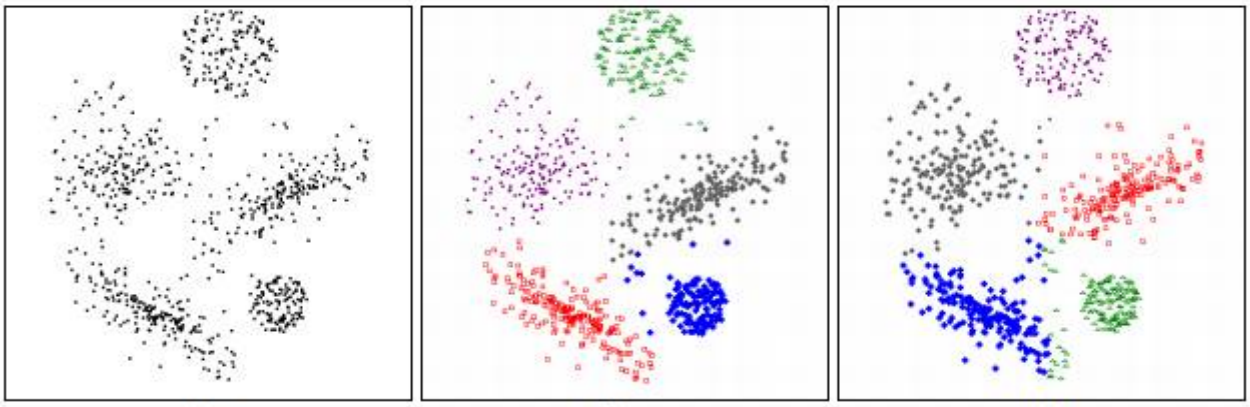

Fig. 1. (a) The information set used in example 2; (b) the clustering result achieved by the GK algorithm in Example 2; (c) the clustering result achieved by the HCK algorithm in Example 1;

V. CONCLUSION

Predicated on the line symmetry distance, an incipient measure LS is then proposed for cluster validation. The simulation results reveal the fascinating observations about the validity measures discussed in this paper. The S validity measure goes wrong if not in one case, at least in the other, and the PC and CE validity measures fail in all the examples. The suggested LS validity evaluate demonstrates that consistency for these and several other examples that are endeavoured. Because the line symmetry distance is elongated from the point symmetry distance, one may interest the performance if the cluster validity measure is predicated on point symmetry. According our simulations, if the clusters with point symmetry structure (e.g. examples 2 and 3), the validity measure predicated on point symmetry should withal perform well for these cases. However, if the geometrical structure of clusters is line symmetry (e.g. example 4), it not a good cull to utilize point symmetry as validity measure.

Albeit these simulations shows that the incipient measure outperforms the other three measures, we optate to accentuate that there are additionally constraints associated with this incipient measure. First, we require surmising that clusters are line symmetrical structures. If the information set does not follow the postulation, the quantification may not work well. Second, for immensely colossal information sets, the tenaciousness of the quantification is very computationally extravagant. The price paid for the flexibility in detecting cluster number is the incrementation of computational involution. The computing involution is O(N2), where sN designates the information number. In fact, an abundance of future work can be done to ameliorate not only the line symmetry distance but withal the LS measure.

REFERENCES

[1] R. C. Dubes and A. K. Jain, Algorithms for Clustering Information, Englewood Cliffs, Prentice Hall, NJ, 1988.
[2] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, Wiley, NY, 2001.
[3] J. Bezdek, Pattern identification with Fuzzy Objective Function Algorithms, Plenum, NY, 1981.
[4] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis-Methods for Classification, Information Analysis and Image Recognition, John Wiley & Sons, Ltd., 1999.
[5] C.-S. Fahn and Y.-C. Lo, "On the clustering of head-related transfer functions used for 3-D sound localization," Journal of Information Science and Engineering, Vol. 19, 2003, pp. 141-157.
[6] P.-C. Wang and J.-J. Leou, "New fuzzy hierarchical clustering algorithms," Journal of Information Science and Engineering, Vol. 9, 1993, pp. 461-489.
[7] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in Proceedings of IEEE Conference on Decision and Control, 1979, pp. 761-766.
[8] R. N. Dave, "Use of the adaptive fuzzy clustering algorithm to detect lines in digital images," Intelligent Robots and Computer Vision VIII, Vol. 1192, 1989, pp. 600-611.
[9] R. N. Dave, "Fuzzy shell-clustering and application to circle detection in digital images," International Journal General Systems, Vol. 16, 1990, pp. 343-355.

ISSN(Online) : 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

[10] Y. Man and I. Gath, "Detection and division of ring-shaped clusters using fuzzy clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, 1994, pp. 855-861.
[11] R. N. Dave and K. Bhaswan, "Adaptive fuzzy c-shells clustering and detection of ellipses," IEEE Transactions on Neural Networks, Vol. 6, No. 5, 1992, pp. 643-662.
[12] F. Höppner, "Fuzzy shell clustering algorithms in image processing: fuzzy c-rectangular and 2-rectangular shells," IEEE Transactions on Fuzzy Systems, Vol. 5, 1997, pp. 599-613.
[13] M. C. Su and Y. C. Liu, "A new approach to clustering information with arbitrary shapes," Pattern Recognition, Vol. 38, 2005, pp. 1887-1901.
[14] J. C. Dunn, "Well separated clusters and optimal fuzzy partitions," Journal Cybern, Vol. 4, 1974, pp. 95-104.
[15] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," Journal of Mathematical Biology, Vol. 1, 1974, pp. 57-71.