DATA CLUSTERING FRAMEWORK BASED ON DENSITY

[1]Vengala Rakshitha, [2]Dr. R. Jegadeesan, [3]P.Balakishan [4]K.Jayasri, [5]B.Anjali, [6]P.Saipriyanka

[1,3,4,5] UG Student, [2,3]Associate professor.-Department of Computer Science and Engineering

[1,2,3,4,5,6]Jyothishmathi Institute of Technology and Science, Karimnagar, India.

Abstract **–**Grouping of information with high measurement and variable densities represents a surprising test to the customary thickness based bunching strategies. As of late, entropy, a numerical proportion of the vulnerability of data, can be utilized to quantify the outskirt level of tests in information space and furthermore select huge highlights in list of capabilities. It was utilized in our new system dependent on the sparsity-density entropy (SDE) to bunch the information with high measurement and variable densities. To begin with, SDE directs great testing for multidimensional information and chooses the agent highlights utilizing sparsity score entropy (SSE). Second, the grouping results and commotions are acquired embracing another thickness variable bunching technique called density entropy (DE). The adequacy and effectiveness of the proposed SDE structure are approved on engineered and genuine informational indexes in correlation with a few grouping calculations. The outcomes demonstrated that the proposed SDE structure simultaneously identified the clamors and handled the information with high measurement and different densities.

Index Terms: Density Entropy, Sparsity-Density Entropy, Sparsity Score Entropy.

I. INTRODUDCTION

DATA agglomeration is one in every of the foremost wide ways in data processing, that has broad applications in pattern recognition, image process, and information compression, among others [1]. Agglomeration algorithms are divided into five categories: partitioned, ranked, grid-based, density based, and model-based. Partitioned agglomeration ways, e.g., K-means[1], K-medoids[11],and Fuzzy C-Means(FCM) , assign the incoming information points into K disjoint subsets, such points at intervals a cluster are a lot of similar than those in numerous clusters. Ranked agglomeration ways embrace each agglomerated and factious methods: agglomerative methods begin with single-point clusters that are in turn incorporate till a particular criterion is reached; divisive methods split an initial cluster of all data points into top-down density grams supported certain criteria, as in Rock [9], Cure [8], and Chameleon [12].

In this paper, not all the input parameters are pre-given however mechanically determined consistent with the characteristics of knowledge. The formula conjointly works on knowledge sets of uniform densities. The most contributions are highlighted as follows:

1) We have a tendency to propose a unique framework, known as the Sparsity Density Entropy (SDE) framework, which might effectively method each low and high-dimensional knowledge.

2) The planned density-based SDE bunch outperforms alternative ways for clusters with variable densities throughevery cluster.

3) The strategy will effectively exclude the world noises supported absolutely the boundary purpose, that is trivial within the density distribution of knowledge points, and additional take away the native noises consistent with the native boundary purpose.

II. RELATED WORK

This section reviews sampling techniques and feature selection methods

1) Statistical Optimal Sample Size

Testing in information handling might be a precondition technique to pre-select examples of prime quality from the underlying data. However, there is an exchange off among precision and productivity of calculations once the exactness improvement winds up immersed at gigantic example sizes. In [7], scientists assessed their live on four gigantic datasets. They found that the resulting tree sizes with SOSS zone unit significantly littler than those with the total size.

2) Feature Selection

As one vital preprocessing step in information agglomeration, feature choice could be a method of selecting a representative and effective set from original options within the high dimensional information house in step with the specified analysis criterion, specified the preserved feature set is most helpful in capturing the intrinsic properties[5]. Feature choice ways are often divided into 3 groups: filter approaches, wrapper approaches, and embedded approaches.

III.RESEARCH METHODOLOGY

Clustering problems on data sets with high dimensions and varying densities make it a landmark and challenge the methods of traditional group analysis. The goal of reducing the dimension is to reduce storage by data compression, to remove the effect of noise features [4], and to extract media features. SDE proposed framework improves the objective function of compactness at home block and distance between clusters.We choose grouping refers to the clusters using

1) The density entropy method 2-D data area or
2) The entropy mode of the capture

A. Basic Definitions of density entropy:

1) Density Metric:

Given the data set D, the distance metric M, $P_i \in D$ (i = 1, 2, ..., N), the density metric of P_i, .The m (P_i) symbol is the sum of the spaces from P_i to the nearest neighbor K (KNN) of P_i, which is referred to as

$$m(pi) = \sum_{g=1}^{k} dist(pj, y) \forall y \in KNN(pi) \quad (2)$$

2) Border Degree

It limits the degree of P_i ($P_i \in D$), referred to as B (P_i), and is the absolute difference between Densities of entropy pand total entropy density. The border points it has minimal effects on the distribution of information.

3) Border set

The border group is defined as a set of points, referred to as B, which depends on the ratio of entropy after removing B in the total universe of all the points in B.

4) The outer-most Border Threshold

A point pi is the outer most border purpose whose border degree is minimal in Bachelor of Divinity. The outer-most border threshold, denoted as O, is outlined by the utmost magnitude relation between m(pi and K N N(pi).

5) The Automatic Border Threshold:

After intervention of the dataset all objects except noise, including data points in B are classified as a set. Then m (P_i), and the maximum value for density measurements for C (j ∈ C) in B is relabeled as the border threshold of C_j, denoted as BT (j).

6) Noise:

A point pi belongs to the Noise1 if $m(p_i) \geq m(q_o)$ and it belongs to Noise2 if $p_i \in C(l)$ and $m(p_i) \geq A(l)$. Noise = Noise1∪ Noise2. In the other word, the density metric of a noise point is larger than the density metric of a border point for each cluster.

Algorithm1: Density Entropy Algorithm[3]
Input: Data set D, with its number of objects *Num*. Output: Noise points Noise, the clustering results Results: 1. According to data set D, determine the value of K, K = int(√*Num*) + 1. 2. According to distance metric, calculate the whole distance metric T. 3. Compute the density metric DM and the KNN of each object. 4. Calculate the border degree of each point B. 5. Compute the border set S. 6. Calculate the outer most border threshold O and Noise1. 7. Perform clustering from the point pi which has the minimum density metric in {B − Noise1} and then pi is labeled as C(l), l =1. 8. We select and extend an unexpended point in C(l) according to the density-reachable concept until all the points in C(l) have been extended. All the objects in C(l) form one cluster. 9. After that, we start a next cluster from a point which has the maximum density metric in {B−Noise1−C(l)}. Go round and begin again until all the objects have been labeled in{B−Noise1}. We obtain C = {C_l|l= 1,2,. . .,m, $C_1 \in C$} and label the objects in S based on C. 10. According to C, labeled S and definition 5, calculate A. 11. According to definition 7 and C, compute Noise2 and obtain Noise and a new clustering results Results = C−Noise2.

B. Sparsity Score Entropy:

In this paper, to handle the 3-d knowledge, we have a tendency to firstly use the above-named SOSS technique to get the sample knowledge set from a given dataset. Consistent with the higher than algorithmic rule, its time quality is O (n2) and area quality is O(n). Our approach supported the full knowledge set would cause higher algorithmic rule quality.

C. Computational Complexity Analysis

The time quality analysis of the SDE is provided in theory by steps. The first step needed to work out K. the entire quality of this step is O (√N). The second step is to get the total distance metric, and therefore the time quality is O (N2).

Algorithm2: Sparsity Score Entropy[6]
Input: Sparsity Score S(r),r = 1, 2, . . ., d, QD = {qD1, qD2, . . ., Lqn} Output: feature subsets F S*, feature weight ω *, similarity entropy distance Sem, dataset QD_FS . 1. Initialize feature subsets F S* and feature weight ω = (1, 1. . . 1) ∈Rd . 2. Separately calculate the entropy of each dimension feature, the total entropy, and the entropy without r-th (r = 1, 2 . . . d) feature. 3. Add the r-th feature to F S* when Emi(r) ≥ 1. 4. Extract the new dataset QD_FS from the dataset QD, and QD_FS contains the corresponding features in FS*. 5. Compute the entropy weights ω * of the selected m features.

IV. EXPERIMENT RESULTS

1) Experimental Settings

To illustrate the potential of the planned SDE framework to handle information with complicated distribution, we feature on many experiments on some artificial datasets and real world datasets. In detail, we tend to compare the performance of our planned American state methodology against variety of progressive clump algorithms, as well as K-means, DBSCAN.

Table 1: Statistics of Experimental Data Sets

Datasets	Source	#Instances	#Features	#Classes
Labro	IRIS	57	16	3
Diabetes	NIDDKD	768	32	2
Soybean	LBTL	683	35	9
Vote	CQA	435	17	2

2) 2-D Data Set

The datasets are challenging for most clustering methods. Most of them could not find the correct range of clusters or have the development of over-fitting. In this paper, Density Entropy might notice the minus distinction among these density metrics. Therefore Density Entropy outperformed the opposite clump strategies on Labro.

The accuracy in Table shows the higher performance of Density Entropy. Information set Soybean that information points with non-uniform densities. There are 134 outliers and 4 clusters, wherever every cluster has its own inner cluster density and form.

3) Multi-dimensional dataset

Five datasets with multi dimensions were used to test our clustering SDE framework in this section. In this experiment, we first used the feature selection methods, to select features and then adopted Density Entropy to partition data sets.

Table 2: Five datasets and their corresponding values of the sample size s

Datasets	#Instances	#s
Labro	57	20
Diabetes	768	255
Soybean	683	121
Vote	435	319

Table 3: Sparsity Score Entropy of all features in Labro data set

Feature	1	2	3	4
SS Entropy	1.260	1.200	1.125	0.980

Table 4: Sparsity Score Entropy of all features in Soybean data set

Feature	1	2	3	4	5	6
SS Entropy	0.948	1.030	0.992	0.932	0.009	1.000

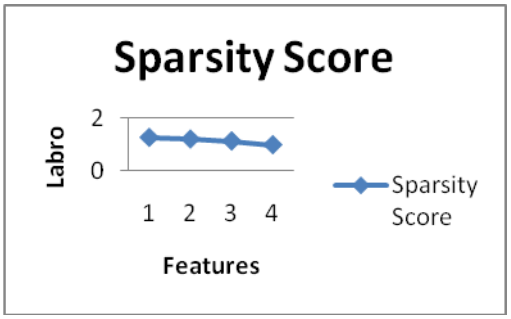

Fig:1. Sparsity score of certain features of Labro

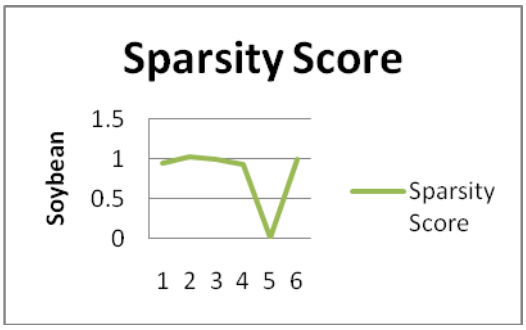

Fig:2. Sparsity score of certain features of Soybean

4) Parameter Sensitivity

There are two key parameters in our planned SDE framework, i.e., K and s. The authors in [7] planned ensemble nearest neighbor classifiers and, within the meanwhile, incontestable the validity of K = √N. Moreover, to gauge the impact of K on the performance of SDE, we have got conducted a collection of experiments on all the datasets. The performance with s outperformed the one with N. Meanwhile, sampling with s shortened the runtime.

Table 6: Comparison by accuracy with SDE and other clustering methods

Datasets	SDE	DBSCAN	K-Mean
Labro	100.0%	98.3%	38.6%
Diabetes	98.6%	72.9%	63.2%
Soybean	94.9%	78.5%	60.8%
Vote	91.8%	70.4%	56.1%

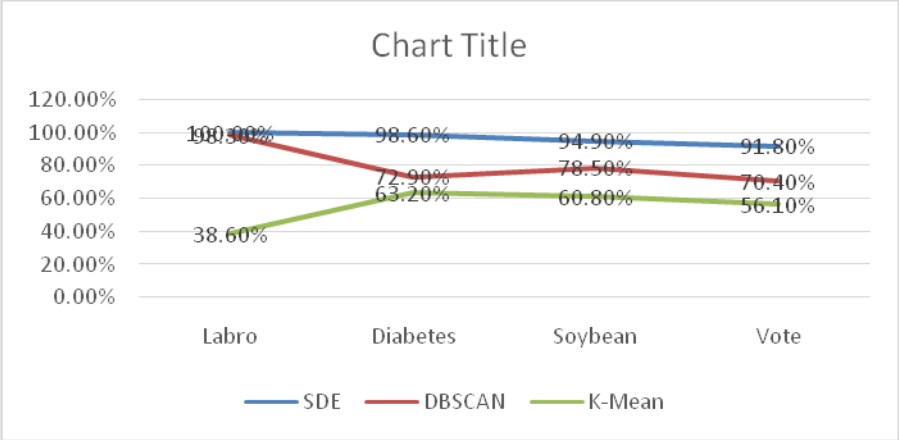

Fig. 4. Sample quality of five datasets with different sizes of data sets by SOSS

5) Performance Evaluation

In order to gauge the run time of our planned SDE framework, we tend to tested on the artificial datasets with completely different sizes and dimensions. The run time was averaged over 10 runs for every dataset with identical parameters. Supported the SDE framework, there are a unit 2 cases regarding the general run time: the agglomeration time solely supported American state; the agglomeration time supported each SSE and DE [10].

The first half suggests that the need of finding the KNN of every rest purpose once more. The second half suggests that precomputing the similarity matrix and the overall entropy while not every feature successively. Thus, we will divide the datasets into several sets and every kernel processes one subset with our planned SDE technique.

Table 7: Accuracy for nine datasets with a range of K

Datasets	3-NN	5-NN	√N-NN
Labro	94.6%	94.8%	100.0%
Diabetes	91.5%	93.2%	98.6%
Soybean	85.7%	87.6%	91.2%
Vote	70.2%	71.4%	73.5%

Table 8: Accuracy and the run time on nine datasets with N and s

Datasets	N	Accuracy	Time	S	Accuracy	Time
Labro	57	93.5	6.2	97	94.9	3.1
Diabetes	768	92.3	11.3	112	91.2	8.2
Soybean	683	83.4	3 8.7	212	85.2	9.5
Vote	435	76.5	44.1s	2850	73.4	23.6

5. CONCLUSION AND FUTURE SCOPE

In this paper, we will in general venture a substitution shifted thickness based programmed grouping frameworkSDE. In SDE, we will in general decide the world external outskirt edge to lead beginning cluster exploitation the DE approach. At the point when toward the begin determinative the cluster scope, we tend to conveyed bunch once more on each local extension with its local external most fringe limit. Through 2-step bunch, we tend to disposed of each the world and local clamors, and moreover grouped the datasets reliable with comparative thickness measurements. Picking partner right limit to select choices on altogether different datasets keeps on being a huge test for the predominant calculations. Also, most bunch calculations need the cluster go as a past; we will in general exclusively must be constrained to set the amount of closest neighbors to survey the thickness measurements of data focuses.

REFERENCES

[1] Jain, A.K., Dubes, R.C. 1998. “Algorithm for Clustering Data”. Printice Hall Englewood cliffs NJ

[2] Han, J., Kamber, M. 2001. “Data Mining: Concepts and Techniques”. Morgan Kaufman

[3] Ester M., Kriegel H.-P., Sander J., Xu X.: “A DensityBased Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, 1996, pp. 226-231.

[4] H. Almuallim and T.G. Dietterich, “Algorithms for Identifying Relevant Features,” Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.

[5] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, “On Feature Selection through Clustering,” Proc. IEEE Fifth Int’l Conf. Data Mining, pp. 581-584, 2005.

[6] S. Berchtold, C. Bohm, D. Keim, and H.-P. Kriegel. “Cost model for nearest neighbor search in high dimensional data space. In Proc. of Symposium on Principles of Database Systems”, Tucson, Arizona,1997.

[7] Ester M., Kriegel H.-P., Xu X.: “Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification”, Proc. 4th Int. Symp. on Large Spatial Databases, Portland, ME, 1995, in: Lecture Notes in Computer Science, Vol. 951, Springer, 1995, pp. 67-82.

[8] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. “CURE: An efficient clustering algorithm for large Databases” In Proc. of 1998 ACM-SIGMOD Int. Conf. on Management of Data, 1998.

[9] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. “ROCK: a robust clustering algorithm for categorical Attributes”. In Proc. of the 15th Int’l Conf. on Data Eng., 1999.

[10] L. Kaufman and P.J. Rousueeuw. (1990) “Finding Groups in Data: an Introduction to Cluster Analysis”, John Wiley & Son

[11] “k-medoids clustering using partitioning around medoids for performing face recognition” Aruna Bhat ,Department of Electrical Engineering, IIT Delhi, Hauz Khas, New Delhi

[12] “CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling” George Karypis Eui-Hong (Sam) Han Vipin Kumar.

[13]Jegadeesan,R.,Sankar Ram M.Naveen Kumar JAN 2013 “Less Cost Any Routing With Energy Cost Optimization” International Journal of Advanced Research in Computer Networking,Wireless and Mobile Communications.Volume-No.1: Page no: Issue-No.1 Impact Factor = 1.5

[14]. Jegadeesan,R.,Sankar Ram, R.Janakiraman September-October 2013
"A Recent Approach to Organise Structured Data in Mobile Environment" R.Jegadeesan et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6) ,Page No. 848-852 ISSN: 0975-9646 Impact Factor:2.93

[15]. Jegadeesan,R., Sankar Ram October -2013 "ENROUTING TECHNICS USING DYNAMIC WIRELESS NETWORKS" International Journal of Asia Pacific Journal of Research Ph.D Research Scholar 1, Supervisor2, VOL -3 Page No: Print-ISSN-2320-5504 impact factor 0.433

[16]. Jegadeesan,R., Sankar Ram, M.S.Tharani (September-October, 2013) "Enhancing File Security by Integrating Steganography Technique in Linux Kernel" Global journal of Engineering,Design & Technology G.J. E.D.T., Vol. 2(5): Page No:9-14 ISSN: 2319 – 7293

[17]. Ramesh,R., Vinoth Kumar,R., and Jegadeesan,R., January 2014 "NTH THIRD PARTY AUDITING FOR DATA INTEGRITY IN CLOUD" Asia Pacific Journal of Research Vol: I Issue XIII, ISSN: 2320-5504, E-ISSN-2347-4793 Vol: I Issue XIII, Page No: Impact Factor:0.433

[18]. Vijayalakshmi, Balika J Chelliah and Jegadeesan,R., February-2014 "SUODY-Preserving Privacy in Sharing Data with Multi-Vendor for Dynamic Groups" Global journal of Engineering,Design & Technology. G.J. E.D.T.,Vol.3(1):43-47 (January-February, 2014) ISSN: 2319 –7293

[19]. Jegadeesan,R.,SankarRam,T.Karpagam March-2014 "Defending wireless network using Randomized Routing process" International Journal of Emerging Research in management and Technology

[20].Jegadeesan,R.,T.Karpagam, Dr.N.Sankar Ram , "Defending Wireless Network using Randomized Routing Process" International journal of Emerging Research in management and Technology ISSN: 2278-9359 (Volume-3, Issue-3) . March 2014

[21]. Jegadeesan,R., Sankar Ram "Defending Wireless Sensor Network using Randomized Routing "International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 9, September 2015 ISSN: 2277 128X Page | 934-938

[22]. Jegadeesan,R., Sankar Ram,N. "Energy-Efficient Wireless Network Communication with Priority Packet Based QoS Scheduling", Asian Journal of Information Technology(AJIT) 15(8): 1396-1404,2016 ISSN: 1682-3915,Medwell Journal,2016 (Annexure-I updated Journal 2016)

[23]. Jegadeesan,R.,Sankar Ram,N. "Energy Consumption Power Aware Data Delivery in Wireless Network", Circuits and Systems, Scientific Research Publisher,2016 (Annexure-I updated Journal 2016)

[24]. Jegadeesan,R., Sankar Ram , and J.Abirmi "Implementing Online Driving License Renewal by Integration of Web Orchestration and Web Choreogrphy" International journal of Advanced Research trends in Engineering and Technology (IJARTET) ISSN:2394-3785 (Volume-5, Issue-1, January 2018

[25]. Pooja,S., Jegadeesan,R., Pavithra,S., and Mounikasri,A., "Identification of Fake Channel Characteristics using Auxiliary Receiver in Wireless Trnsmission" International journal for Scientific Research and Development (IJSRD) ISSN (Online):2321-0613 (Volume-6, Issue-1, Page No. 607-613, April 2018

[26]. Sangeetha,R., Jegadeesan,R., Ramya,P., and Vennila.,G "Health Monitoring System Using Internet of Things" International journal of Engineering Research and Advanced Technology (IJERAT) ISSN :2454-6135 (Volume-4, Issue-3, Page No. 607-613, March 2018.