iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IX **Month of publication:** September 2019

DOI: http://doi.org/10.22214/ijraset.2019.9106

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue IX, Sep 2019- Available at www.ijraset.com

Ensemble Mixed Breed Deep Clustering Algorithm for Complex Datasets

S. Nagarjuna Reddy[1], S. Sai Satyanarayana Reddy[2], M. Babu Reddy[3]
[1]*Research Scholar, Dept of CSE, JNTUK, Kakinada, India*
[2]*Principal, Vardhaman College of Engineering, Hyderabad, India*
[3]*Associate Professor, Dept of CSE, Krishna University, Machilipatnam, India*

***Abstract:** Clustering is process of having similar items at one place. Many datasets are available for the research in clustering. Machine learning (ML) and deep learning (DL) are two latest domains that to improve the clustering techniques. From past decade so many applications are developed in clustering for pattern recognition, speech and other prediction type of algorithms. According to the latest research, deep clustering algorithms can be used to learn better representations of the data. In this paper, the Ensemble Mixed Breed Deep Clustering Algorithm (EMBDCA) which is adopted various deep learning algorithms for improving the performance. For the training, Information Maximizing Self-Augmented Training (IMSAT) is utilized. This will improve the accuracy especially for the datasets such as mushroom and MIST dataset. The parameters sensitivity, specificity and quality of clusters are also improved.*

***Keywords:** Clustering, IMSAT, Deep Learning, Machine Learning*

I. INTRODUCTION

Clustering algorithms will be reliant on the variety of the input information provided, such that various datasets could require diverse similarity measures and distinctive separation methods. Therefore, dimensionality decrease and portrayal learning have been widely utilized nearby grouping, so as to outline input information into a component space where partition is simpler regarding the issue's unique circumstance. Utilizing deep neural networks (DNNs), it is conceivable to learn non-direct mappings permitting to change the information into all the more clustering-friendly representations. Clustering, one of the principal territories in AI targets ordering unlabeled information into gatherings (clusters). A promising course in profound learning exploration is to learn portrayals and at the same time find cluster structure in unlabeled information by enhancing a discriminative misfortune work. Deep Embedded Clustering (DEC) [2] epitomizes this profession and speaks to, as far as we could possibly know, the cutting edge. DEC depends on an improvement methodology wherein a neural system is pertained by methods for an auto encoder and afterward tweaked by mutually upgrading bunch centroids in yield space and the fundamental component portrayal. Another model is [3], where the creators propose a joint enhancement for dimensionality decrease utilizing a neural system and k-implies grouping. Different ways to deal with single deep learning dependent on antagonistic systems have as of late been proposed [4]. These methodologies are distinctive in soul however can likewise be utilized for clustering.

II. RELATED WORK

Clustering techniques that consider the linkage between information focuses, generally known as various leveled strategies, can be subdivided into two gatherings: agglomerative and troublesome. In an agglomerative progressive grouping calculation, at first, each item has a place with a particular individual bunch. At that point, after progressive emphases, bunches are converged until stop conditions are come to. Then again, a troublesome various leveled bunching strategy begins with all articles in a solitary group and, after progressive cycles, objects are isolated into bunches. There are two principle bundles in the R language that give schedules to performing various leveled grouping, they are the details and bunch. Here we consider the agnes routine from the bunch bundle which actualizes the calculation proposed. Four understood linkage criteria are accessible in agnes, specifically single linkage, complete linkage, Ward's strategy, and weighted normal linkage. Clustering is a great data preparing issue, especially significant in AI [5, 6, 7, 8, 9]. Endless methodologies exist for clustering, with mean move, k-means and desire augmentation calculations [10], being probably the most outstanding ones. In the most recent decade, unearthly grouping assumed a noticeable job in the field, see for example [11-15]. Spectral clustering misuses the range of closeness lattices to parcel input information. In spite of the fact that these strategies have exhibited great execution in complex issues, they experience the ill effects of absence of adaptability as for the quantity of info information focuses; cubic computational multifaceted nature for Eigen solvers and quadratic intricacy as far as memory occupation. Endeavours to take care of these issues have been made by planning approximations or utilizing various advancement methods.

III. DATASET DESCRIPTION

A. *Mushroom Dataset*

It is one of the most popular data set contains total of 8124 training occurrences, among them each occurrence represents single mushroom. The first attribute is the target variable containing the label names used to recognize whether the mushroom is belongs to edible or poisonous group.

B. *MNIST Dataset*

This data set contains set of handwritten digits, contains total of 60,000 training occurrences available and 10000 examples will be considered as test set. All the hand written digits are normalized.

IV. INFORMATION MAXIMIZING SELF-AUGMENTED TRAINING (IMSAT)

It represents the data using information maximization between input and cluster assignment. It proposes Self Augmentation Training, which penalizes representation dissimilarity between the original data points and augmented ones, T(x).

$$\mathcal{R}_{\mathrm{SAT}}(\theta; x, T(x)) = -\sum_{m=1}^{M}\sum_{y_m=0}^{V_m-1} p_{\hat{\theta}}(y_m|x)\log p_{\theta}(y_m|T(x))$$

It combines mutual information constraint along with SAT scheme to define objective function as:

$$\min \mathcal{R}_{\mathrm{SAT}}(\theta; T) - \lambda[H(Y) - H(Y|X)]$$

A. *The Ensemble Clustering Algorithm*

Input: $H^{(1)}$... $H^{(r)}$, r basic partitions
L represented as layer count
p: Considered Noise Level
K: Cluster size
Output: optimal H^{*}

1) Initially Construct the Binary Matrix (B)
2) Obtain Mapping Matrix W by applying layered stacked EMDCA by taking noise level p.
3) : Apply K-means algorithm on BW^{T} to obtain H^{*}

V. PERFORMANCE EVOLUTION

Based on the proposed algorithm to analyze the performance of the system use various measures Accuracy, purity of the cluster etc., the fundamental count values available in the confusion matrix such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are used to estimate the measures

A. *False Positive Rate (FPR)*

The fractional amount cases it will be classified as correct, but it is wrong.

$$FPR = \frac{FP}{FP + TN}$$

B. *False Negative Rate (FNR)*

The fractional amount of cases it will be classified as wrong, but it is correct.

$$FNR = \frac{FN}{FN + TN}$$

C. *Sensitivity*

The fractional amount of actual positive values which are successfully identified.

$$Sesitivity = \frac{No.of\ TP}{No.of\ TP + No.of\ TN}$$

D. Specificity

The fractional amount of actual negative values which are successfully identified

$$Specificity = \frac{No.of\ TN}{No.of\ TN + No.of\ FP}$$

Accuracy: This will calculate the overall accuracy of the clusters.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

E. Normalized Mutual Information (NMI)

NMI is also called data theoretic parameter which calculates the mutual data between the ground truth labels and cluster assignments. This is normalized based on average of entropy of both ground labels and the cluster assignments. The formula to calculate the NMI score for the mushroom and MIST datasets is.

$$NMI(Y,C) = \frac{I(Y,C)}{\frac{1}{2}[H(Y) + H(C)]}$$

The implementation is done with IMSAT is used for the training of the dataset to improve the performance. By using Java programming language which performs better to get the results. EMBDCA utilized the mushroom dataset and the MNIST database of handwritten digit datasets is used. The performance of the four algorithms is compared.

VI. RESULTS OF SEGMENTATION EVALUATION

The mushroom data set represented in the following figure:2 it contains set of training examples, it is available in comma separated value format.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1	cap-shape	cap-surfac	cap-color	bruises	odor	gill-attach	gill-spacin	gill-size	gill-color	stalk-shap	stalk-root	stalk-surfa	stalk-surfa	stalk-colo	stalk-colo	veil-type	veil-color	ring-numb	ring-type	spore-prir	pop
2	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s
3	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	n
4	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	n
5	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s
6	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a
7	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	n
8	b	s	w	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	n
9	b	y	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	s
10	x	y	w	t	p	f	c	n	p	e	e	s	s	w	w	p	w	o	p	k	v
11	b	s	y	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	s
12	x	y	y	t	l	f	c	b	g	e	c	s	s	w	w	p	w	o	p	n	n
13	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	s
14	b	s	y	t	a	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n	s
15	x	y	w	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	v
16	x	f	n	f	n	f	w	b	n	t	e	s	f	w	w	p	w	o	e	k	a
17	s	f	g	f	n	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	y
18	f	f	w	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a
19	x	s	n	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s
20	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n	s
21	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	s
22	b	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	s
23	x	y	n	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n	v

Figure: 2, Sample Mushroom dataset

The following table discuss about the results compared with EM-GMM, Mixed Breed.

Cluster-1	Accuracy	Sensitivity	Specificity	No of records in this cluster	Quality	Purity	NMI
EM-GMM	0.99768	0.99	0.96	0.51	0.87	0.67	-
Mixed Breed	0.99987	0.99987	0.98	0.61	0.97	0.96	-
EMBDCA	0.9999	0.999	0.99	0.87	0.98	0.98	0.99

Table: 1, In Cluster-1 performance for the mushrooms are belongs to edible.

Cluster-2	Accuracy	Sensitivity	Specificity	No of records in this cluster	Quality	Purity	NMI
EM-GMM Clustering	0.9976	0.9943	0.96	0.48	0.87	0.89	-
Mixed Breed	0.9997	0.9994	0.98	0.49	0.97	0.96	-
EMBDCA	0.9999	0.999	0.99	0.68	0.98	0.98	0.99

Table: 2, in cluster-2 all the mushrooms are belongs to poisonous.

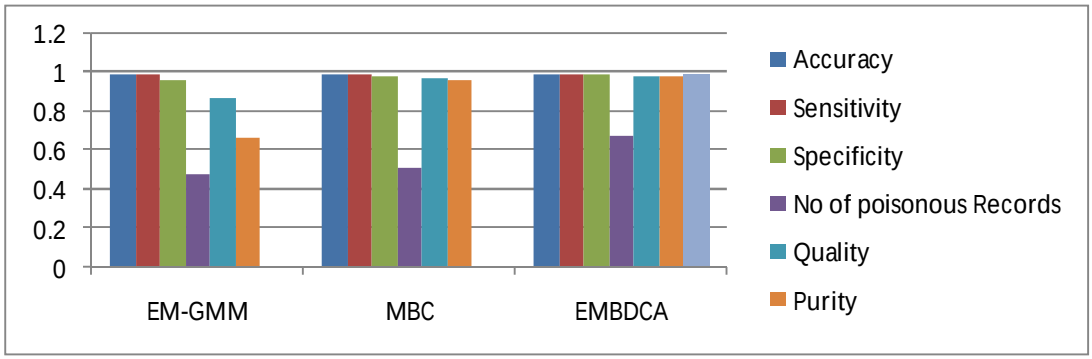


Figure: 3 Performance graph representation for edible (Cluster-1) records in mushroom dataset.

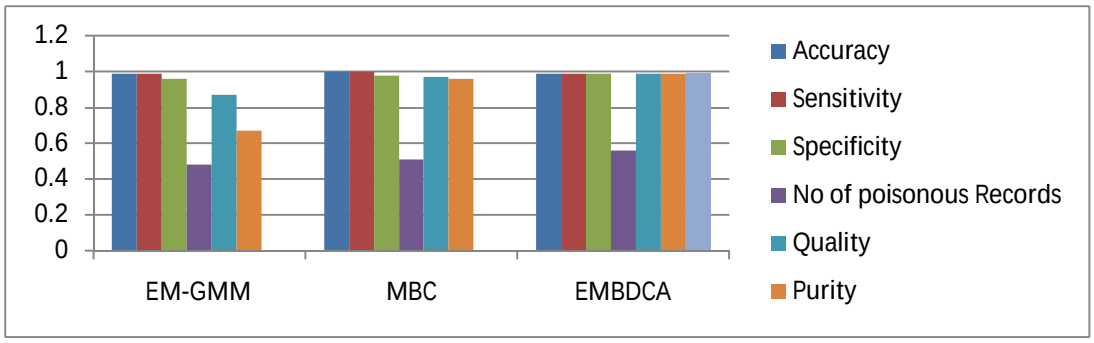


Figure: 4 Performance graph representation for poisonous (Cluster-2) records in mushroom dataset.

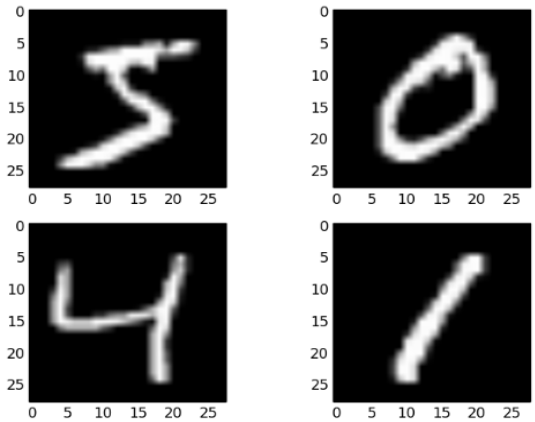


Figure: 5 Examples of MNIST dataset

Evolution results on MIST hand written dataset

	Accuracy	Sensitivity	Specificity	No of records in this cluster
Cluster-1 (Number 0)	0.99	0.97	0.98	0.432
Cluster-2 (Number 1)	0.99	0.98	0.97	0.453
Cluster-3 (Number 2)	0.99	0.99	0.98	0.543
Cluster-4 (Number 3)	0.99	0.94	0.95	0.654
Cluster-5 (Number 4)	0.99	0.98	0.98	0.453
Cluster-6 (Number 5)	0.99	0.9998	0.96	0.675
Cluster-7 (Number 6)	0.99	0.9967	0.97	0.567
Cluster-8 (Number 7)	0.98	0.9786	0.96	0.568
Cluster-9 (Number 8)	0.98	0.9909	0.96	0.5654
Cluster-10 (Number 9)	0.97	0.97	0.97	0.5654

Table: 3 Show the performance of the clusters for handwritten dataset 0-9

The EMBDCA outcomes of clustering of data are presented by our proposed work. The comparison between to our EM-GMM clustering, MBC, EMBDCA in this comparison our proposed technique will give very high accuracy values for clustering of data. Among the two existing clustering algorithm the EMBDCA performs well based on the parameters such as sensitivity, specificity, accuracy, purity, quality and NMI.

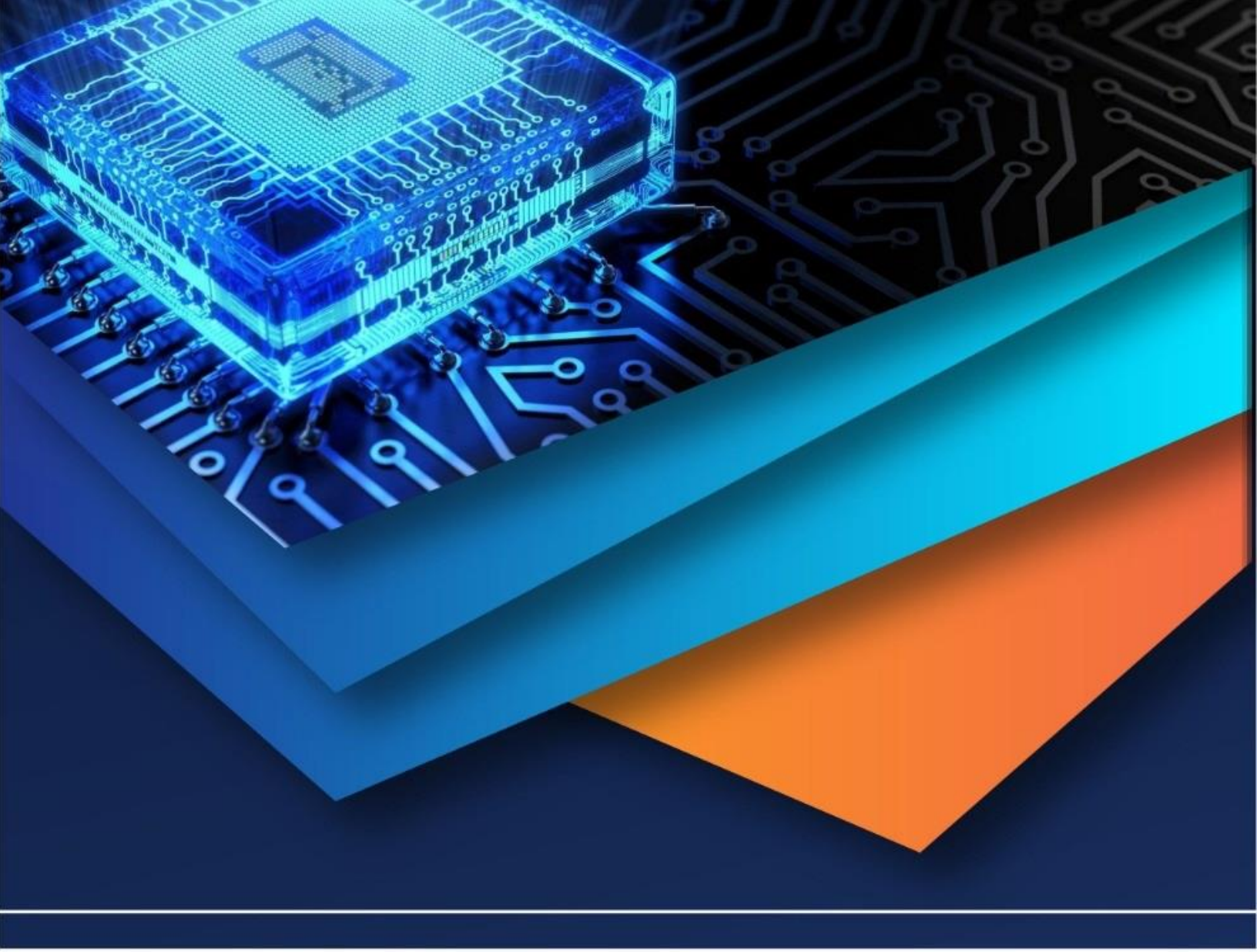
VII. CONCLUSION

In this paper, the EMBDCA is the ensemble approach which performs well compare with the other existing clustering algorithms. IMSAT is utilized to make the clustering process easy and improves the accuracy and NMI score by using the training algorithm. After training algorithm, the proposed EMBDCA process the data original data and forms the clusters based on various parameters discussed in above section. This is also considered the complexity and compatibility of the algorithm weather it is fit for the processing of mushroom and MINST datasets. In future, EMBDCA adopted with big data algorithms to process the mega bytes data for clustering.

REFERENCES

[1] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (4) (2007) 395–416.
[2] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: Proceedings of the 33rd International Conference on International Conference on Machine Learning, Vol. 48, JMLR.org, 2016, pp. 478–487.
[3] B. Yang, X. Fu, N. D. Sidiropoulos, M. Hong, Towards k-means-friendly spaces: Simultaneous deep learning and clustering, arXiv preprint arXiv:1610.04794.
[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[5] A. K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognition Letters 31 (8) (2010) 651–666.
[6] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.
[7] F. M. Bianchi, L. Livi, A. Rizzi, Two density-based k-means initialization algorithms for non-metric data clustering, Pattern Analysis and Applications 19 (3) (2016) 745–763. doi:10.1007/s10044-014-0440-4.
[8] F. Nie, L. Tian, X. Li, Multiview clustering via adaptively weighted procrustes, in: Proceedings of the 24th ACM SIGKDD. International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 2022–2030.
[9] J. N. Myhre, K. Ø. Mikalsen, S. Løkse, R. Jenssen, Robust clustering using a knn mode seeking ensemble, Pattern Recognition 76 (2018) 491–505.
[10] C. C. Aggarwal, C. K. Reddy, Data Clustering: Algorithms and Applications, CRC Press, Boca Raton, Florida, US, 2013.

[11] R. Jenssen, Kernel entropy component analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (5) (2010) 847–860. doi:10.1109/TPAMI.2009.100.
[12] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, C. Zhang, Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering, IEEE Transactions on Neural Networks 22 (11) (2011) 1796–1808.
[13] Y. Yang, D. Xu, F. Nie, S. Yan, Y. Zhuang, Image clustering using local discriminant models and global integration, IEEE Transactions on Image Processing 19 (10) (2010) 2761–2773.
[14] J. Yang, D. Parikh, D. Batra, Joint unsupervised learning of deep representations and image clusters, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5147–5156.
[15] X.-L. Zhang, Multilayer bootstrap networks, Neural Networks 103 (2018) 29–43.
[16] Hu, W., Miyato, T., Tokui, S., Matsumoto, E. and Sugiyama, M., "Learning discrete representations via information maximizing self-augmented training", 2017. arXiv preprint arXiv:1702.08720.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY